June 2022

ReCent
Actuarial News

Generate insights traditional methods fail to find.

hr | bluebox: Practical insights to machine learning

There is an industry-wide understanding that advanced data analytical methods will shape the way we conduct business in the future. However, many players in the (re)insurance world lack practical examples on how to reap the fruits of this expertise. At Hannover Re, our hr | bluebox service uses machine learning algorithms applied by our skilled data scientists to detect the drivers of early lapses in the portfolios of our clients.

What do we do?

With our free-of-charge service, we aim to identify and characterise portfolio segments with notable early lapse behaviour. This will enable us to isolate likely-to-lapse as well as unlikely-to-lapse policyholders from the rest of the portfolio and define simple rules to recognise these policyholders in the future.

Such a rule could look like: "For policyholders living in region A, C or F with occupation code 1, 2, 3, 4 or 5, we expect an average early lapse rate of 83%". Together with our clients, we will use these insights to decrease the early lapse rate. A typical business implementation might be that our results lead to a more tailored marketing campaign by focussing on the unlikely-to-lapse policyholders. This way, fewer acquisition costs are spent on policyholders who lapse before they can become profitable.

The definition of an "early" lapse varies from client to client, often depending on what causes the most financial damage, e.g. lapses within the first 6 months or lapses within the first year.

Reasons for using machine learning

Machine learning becomes particularly advantageous when one considers the high number of possible rules for a typical data set.

To illustrate this point, let us have a brief look at the possible numbers of combinations for one single rule: Given only two explanatory variables, such as occupation code and region of the policyholder, with 10 distinct values each (occupation code A – J and region 1 – 10), we end up with 1,046,528 non-trivial ways of defining a rule.

Considering that a typical data set can contain up to 100 different explanatory variables, it is impossible to evaluate all possible rules manually. In contrast, the "Classification and Regression Trees" (CART) algorithm works efficiently with high dimensional data. Additionally, CART inherently aims at isolating low- and high-lapsing policyholders in the portfolio. This is done by systematically constructing and evaluating numerous rules to eventually narrow it down to a few essential ones.

These rules are based on and constrained by the explanatory variables available for each individual portfolio. By carefully evaluating the resulting rules and corresponding segments, we generate business insights ready for implementation.

How does CART work?

The CART algorithm aims at maximising the 'purity', in terms of the lapsing behaviour, of the portfolio segments. A segment $\boldsymbol{A}$ in which all policies are lapsed, or all policies are non-lapsed, is 100% pure and is assigned a so-called impurity value of $I(\boldsymbol{A}) = 0$. Conversely, a segment in which exactly half of the policies are lapsed is 100% impure. A commonly used impurity measure is the Gini impurity, defined as:

$$I(\boldsymbol{A}) = 2 \cdot p \cdot (1 - p),$$

where p denotes the share of lapsed policies within segment $\boldsymbol{A}$. The aim of the CART algorithm is now to partition the portfolio into segments such that the weighted sum of all segments' impurities is minimal.

CART proceeds in a greedy (stepwise) fashion: First, the portfolio is split into only two segments based on a simple pair of rules, e.g. "The policyholder lives in region 1 or 3" versus "The policyholder does not live in region 1 or 3".

This can be visualised in the form of an (inverted) tree, where the 'root' $\boldsymbol{A}_0$ at the top represents the complete portfolio and the 'branches' $\boldsymbol{A}_L$ (left) and $\boldsymbol{A}_R$ (right) represent the two segments resulting from this first split. The splitting condition is chosen from all possible options (e.g. region 1 vs. not, region 2 vs. not, region 1 or 2 vs. not, occupation class A vs. not, etc.) such that it results in the highest decrease in impurity, as expressed in the following formula:

$$\Delta I = n_0 \cdot I(\boldsymbol{A}_0) \; - \; [\, n_L \cdot I(\boldsymbol{A}_L) \; + \; n_R \cdot I(\boldsymbol{A}_R) \,]$$

Here, $I(\boldsymbol{A}_0)$, $I(\boldsymbol{A}_L)$ and $I(\boldsymbol{A}_R)$ are the impurities in $\boldsymbol{A}_0$, $\boldsymbol{A}_L$ and $\boldsymbol{A}_R$, respectively. They are weighted by the sample sizes n_0, n_L and n_R.

After the first split has been established, the CART algorithm iteratively splits the portfolio into smaller segments, thereby growing the tree larger and larger. At each split, the algorithm systematically tests all reasonable conditions that may be added to one of the existing rules (e.g. "The policyholder lives in region 1 or 3 AND the insurance sum is larger than EUR 3,000") and selects the condition that results in the highest decrease in impurity as before.

To clarify, consider a simple example data set with 10 policies, as illustrated in Figure 1. There are two explanatory variables characterising each policy: region (circle or diamond) and occupation class (blue or cyan). Lapsed policies are marked with a red cross. The Gini impurity of the portfolio as a whole (root at the top of the tree) is:

$$I(\boldsymbol{A}_{\text{portfolio}}) = 2 \cdot p \cdot (1 - p) = 2 \cdot 0.5 \cdot 0.5 = 0.5.$$

The CART algorithm has two options for the first split: It could either split between region 1 and region 2, or between occupation class A and occupation class B. The decrease in impurity for the first option is:

$$\begin{aligned} \Delta I^{\text{region}} &= n_{\text{portfolio}} \cdot I(\boldsymbol{A}_{\text{portfolio}}) \\ &\quad - [\, n_1 \cdot I(\boldsymbol{A}_1) \; + \; n_2 \cdot I(\boldsymbol{A}_2) \,] \\ &= 10 \cdot 0.5 - [6 \cdot 2 \cdot 1/3 \cdot 2/3 + 4 \cdot 2 \cdot 3/4 \cdot 1/4] \\ &= 0.83, \end{aligned}$$

while for the second option, it would be only

$$\begin{aligned} \Delta I^{\text{occ}} &= n_{\text{portfolio}} \cdot I(\boldsymbol{A}_{\text{portfolio}}) - [\, n_{\text{A}} \cdot I(\boldsymbol{A}_{\text{A}}) \; + \; n_{\text{B}} \cdot I(\boldsymbol{A}_{\text{B}}) \,] \\ &= 10 \cdot 0.5 \\ &\quad - [5 \cdot 2 \cdot 3/5 \cdot 2/5 + 5 \cdot 2 \cdot 2/5 \cdot 3/5] \\ &= 0.20. \end{aligned}$$

Hence, region is chosen for the first split. The algorithm then evaluates whether a second split within the policies from region 1 is worthwhile in terms of the decrease in impurity. It turns out that the decrease in impurity would be 0; hence, no further split is made on the left-hand side of the tree. Conversely, within the policies from region 2, splitting between occupation classes A and B yields a non-zero decrease in impurity of 0.52.

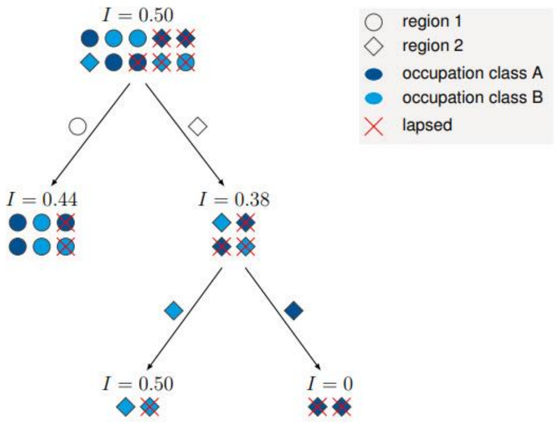

Fig. 1: Example classification tree

The myth of the fully automated workflow

We should take the opportunity here to address a common misunderstanding of machine learning. Namely, that an analysis requires little to no manual work since results will be generated (almost) fully automated. Unfortunately, this is often far away from the truth.

Our typical project cycle starts with the preparation and cleaning of the raw data which was submitted by the client. This includes quality checks, detection of missing values and more.

Even though we developed packages to assist us in the process, many data preparation steps need to be done manually to ensure the quality of the results. Apart from data checks, we also construct new variables for the analysis. These new variables might be based on the available information in the data set but might also be enriched by external data sources (such as socio-economic status based on the zip-codes).

Typically, this process consumes a large share of the project cycle and must not be underestimated. However, for many of our clients, being provided with this cleaned and enriched data set is already a great support.

Once the data is in the correct format, we fit the model. This process is mostly automated and does not take up too much time. Having a fully-fitted model leaves us with two follow-ups:

- First, checking the quality of the identified segments and ensuring that the results can be kept for future use.
- Second, translating the results into tangible actions from a business perspective.

These two steps distinguish our hr | bluebox solution from other services in the market.

Value added by hr | bluebox: Reliable lapse predictions

Ensuring the quality of our results has many different angles. Our main goal is to make the identified segments and rules as reliable as possible.

One threat towards reliability are chance finds: Even if there is no systematic relationship between lapses and the explanatory variables, the CART algorithm is able to detect portfolio segments in which the lapse rate is slightly above or below average due to chance alone – a classic case of overfitting. We thus need to ensure that there is more to our segments than noise.

Among the tools we use to this end is the so-called funnel plot in Figure 2. The labelled points in the plot represent the portfolio segments identified by our analysis, characterised by how much percent of the portfolio they cover (x-axis) and their lapse rate (y-axis). These are compared to the funnel (grey area), which shows for each coverage value, the range of lapse rates we would expect if lapsing was completely random, i.e. independent of all the explanatory variables.

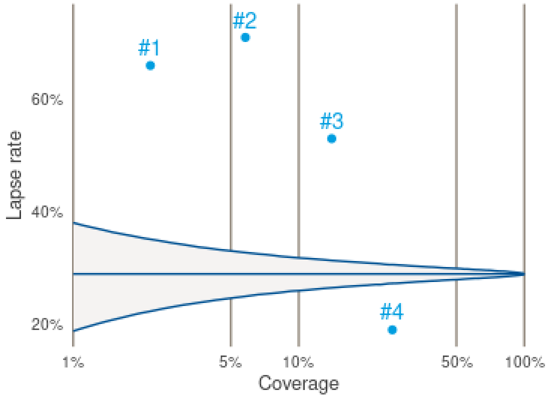

Fig. 2: Funnel plot

Please note that this plot is based on an artificial data set, not the data in the above example. We see that the lower the coverage, the higher the range of lapse rates that can occur by pure coincidence, which meets the expectation.

More importantly, we see a significant gap between the funnel line and the selected points. Analysing point #1 (portfolio segment #1), we see that for the coverage of 2.2% the funnel range lies within 27.5% to 40.6%. The distance between the funnel line and the lapse rate in the selected segment, which is 76% for segment #1, gives us an important indication of the reliability. In this case, since the gap between the point and the funnel line is huge, we can be relatively sure that our segment is not a chance find.

Once the lapse segments are validated, we check if the effect we see of one variable is actually driven by another one. To do this, we compare multiple models for which we repeatedly omit the most important variables from the data set. If as a result an effect in a comparable magnitude is driven by a different variable, we can discuss and clarify this with the client.

One implication could be that we can substitute the variable in the rule with a different one that is easier to implement from a business perspective. However, to visualise this kind of change, we need to interactively visualise the impact of changing the rules submitted by the client.

Adjusting rules interactively: Data visualisation

To explain a change of rules, we developed a dashboard which allows us to interactively visualise the meaning of the derived segments.

Furthermore, for each segment we can change the conditions, as well as exclude or change certain variables, and even add completely new rules. The dashboard then shows the adjusted lapse rate and the coverage of the portfolio. This tool helps us to develop a mutual understanding of the analysis conducted but also to facilitate and optimise the implementation of the rules identified.

Summary

Using machine learning facilitates our analysis and helps us detect complex drivers of early lapses. However, the workflow can never be fully automated and implementing results in a real-world use case requires much more than simply fitting a model. But with our tools and strategies we can generate insights traditional methods fail to find.

Author

Lukas Herrmann
L&H - Data Analytics
Tel. +49 511 5604-2630
lukas.herrmann@hannover-re.com

Follow us on **LinkedIn** to keep up to date with the latest Life & Health news.

The information provided in this document does in no way whatsoever constitute legal, accounting, tax or other professional advice. While Hannover Rück SE has endeavoured to include in this document information it believes to be reliable, complete and up-to-date, the company does not make any representation or warranty, express or implied, as to the accuracy, completeness or updated status of such information. Therefore, in no case whatsoever will Hannover Rück SE and its affiliated companies or directors, officers or employees be liable to anyone for any decision made or action taken in conjunction with the information in this document or for any related damages.