MILLIMAN WHITE PAPER

Claims cost analysis in German private health insurance

Methodology and results

Alexander Krauskopf, DAV, CERA
Katrin Höllering, MSc
Alexandra Pflumm, DAV

This paper considers the use of machine learning algorithms in German health insurance claims cost analyses. The goal is to find a good prediction of future claims cost for chronic diseases. In the first section we introduce the German healthcare system. Our analysis is focused primarily on private health insurance, which is only one part of the German health insurance system. We explain how such an analysis can be executed, which methods are applicable and how the results can be used for disease management.

Overview of the German health system

The German healthcare system is a dual health insurance system with a statutory health insurance (SHI) pillar, complemented by private health insurance (PHI). By law, all German residents are required to have health insurance.

The SHI is part of the German social security system. In general, every person in Germany is insured in the SHI, which means each person is a member of a statutory "health fund" (there are approximately 110 in Germany). The SHI benefit plan is regulated by the Social Security Code (SGB) and can be changed by the government if necessary (for example to reduce the overall claims costs). SHI is financed as a pay-as-you-go system, with no risk-adjusted elements for the individual premium. Every member pays an income-dependent premium (half sponsored by the employer or by the German pension fund for retirees). Family members without their own incomes are also insured, but do not have to pay any premiums. All premiums—together with subsidies from taxes—are collected by the SHI (in the health funds), which then finances all covered treatment, including inpatient, outpatient, dental, and daily allowances in case of disability. Within the SHI there is a risk equalisation process (Morbi-RSA) to equalise claims between different health funds. SHI funds pay physicians directly (through the association of SHI physicians).

Under specific circumstances (e.g., income exceeds a given threshold, self-employed, civil servants), German citizens can opt out of the SHI system and change to a full-coverage tariff in the PHI system (substitution of SHI). In contrast to the SHI, in the PHI system there are many different tariffs with a variation of benefit plans (comprehensive or necessary coverage, high or low prices, with or without deductibles). Premium calculations are based on lifetime pricing, which means that the premium paid by the individual is level over that person's whole lifetime (apart from premium adjustments due to medical inflation, mortality rates or change in actuarial interest rates) and depends only on the insured's age at policy inception. As a consequence, an aging reserve has to be built up in the younger years, which is used for financing the higher claims cost in the older ages. There is also a medical underwriting assessment at the inception of a PHI contract. Benefits in the PHI system are much more flexible than in the SHI, especially in terms of coverage for new methods of treatments or new prescription drugs. The billing process is also different from that of the SHI. Physicians send an invoice to the patient, who has to pay out of pocket and then apply for reimbursement by the PHI company. Therefore, there is usually no direct relationship between physicians and PHI companies.

Approximately 90% of the German population are insured in the SHI, the other 10% are insured in the PHI as a substitution. Within the PHI system, almost 50% are public sector employees.

The standard benefit plan in SHI is limited in some areas (e.g., reimbursement of medical aids or dental prothesis). Therefore, nearly 30% of people insured by SHI prefer to enhance their coverage with supplementary insurance to cover the deficiencies in the SHI.

In the following analysis, we only consider full-coverage substitutive private health insurance.

Motivation for an analysis of claims cost in PHI

A special feature of the German PHI system is that the insurance company is not allowed to cancel a member's contract; there is a lifetime guarantee for the policyholder. As a consequence, the insurer is obliged to review the premium annually and to adjust the premium if necessary (this "premium adjustment clause" is regulated by the Insurance Contract Act and the Insurance Supervision Act).

Simulations from the German Association of Actuaries (DAV) (1) show that medical inflation is expected to be one of the main drivers for the increase of premiums in the PHI system in future decades.

Therefore, the assessment and management of the claims cost in the PHI (especially for chronic diseases) is very important in terms of the sustainability of the policyholders, public reputation and competitiveness for the days ahead.

Case study

INPUT AND PREPARATION

The aim of our case study is to estimate the probability of a diabetes type 2 diagnosis within the next 12 months by using the individual three-year diagnosis histories of PHI policyholders. We used a representative data set for the substitutive PHI from a German private health insurance company, which includes personal information like age, gender and current premium tariff, as well as the diagnosis history from 2010 to 2018. The diagnoses are given as ICD-10-GM codes.

To minimise the number of diagnoses for our analysis (ICD-10-GM has more than 16,000 diagnosis codes) we only use the first three numbers of the ICD-10-GM code. Therefore, for our predictive modeling there are 245 descriptive variables, consisting of 241 ICD codes and information about gender, age and tariff. The target variable is a categorical variable that indicates a diabetes diagnosis (0/1).

All contracts that do not have a three-year data history are removed.

The original data set contains 72,371 insured people from 79 distinct tariffs with 1,884,505 diagnoses (before mapping to the first three digits of ICD-10_GM). We used a subset of the data with full coverage tariffs, which includes 21,101 people and 576,742 diagnoses after validation.

For every policyholder in the data set a single line is created as shown in the table in Figure 1.

FIGURE 1: EXCERPT FROM THE DATA MATRIX FOR THE XG BOOST

Sex	Diabetes	age_diab_2018	current_tariff	A00	A15	A20	…
1	0	60	xxxx	0	0	0	…
1	1	61	yyyy	2	0	0	…
…	…	…	…	…	…	…	…

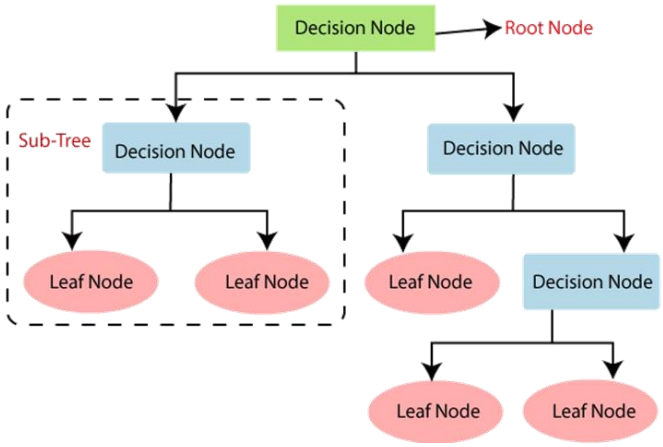
PREDICTION ALGORITHM

We discuss in this section the various prediction algorithm methodology options and their advantages and disadvantages in the context of our case study.

Decision tree

A decision tree is a flowchart-like structure in which each internal node represents a test on an attribute (e.g., whether the policyholder is male or female). Each branch represents the outcome of the test, and each leaf node represents a class label. The paths from root to leaf represent classification rules.

FIGURE 2: DECISION TREE (2)



Decision trees can be used for regression as well as for classification problems. For a regression problem it computes a continuous value (e.g., the probability of a possible outcome) and for a classification problem it returns a natural number (e.g., 0, 1, 2, …, k).

The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

Some advantages of decision trees are:

- Simple to understand and to interpret. Trees can be visualised.
- Requires little data preparation. Other techniques often require data normalisation, dummy variables need to be created and blank values to be removed. Note, however, that this module does not support missing values.
- The cost of using the tree (i.e., predicting data) is logarithmic in the number of data points used to train the tree.
- Able to handle both numerical and categorical data. Other techniques are usually specialised in analysing data sets that have only one type of variable.
- Able to handle multiple-output problems.
- Uses a white box model. If a given situation is observable in a model, the explanation for the condition is easily explained by Boolean logic. By contrast, in a black box model (e.g., in an artificial neural network), results may be more difficult to interpret.
- Possible to validate a model using statistical tests. That makes it possible to test the reliability of the model.

The disadvantages of decision trees include:

- Decision-tree learners can create overly complex trees that do not generalise the data well. This is called overfitting. Mechanisms such as pruning, setting the minimum number of samples required at a leaf node or setting the maximum depth of the tree are necessary to avoid this problem.
- Decision trees can be unstable because small variations in the data might result in completely different trees being generated.
- The problem of training an optimal decision tree is known to be NP-complete[1] under several aspects of optimality and even for simple concepts. Consequently, practical decision-tree learning algorithms are based on heuristic algorithms such as the greedy algorithm, where locally optimal decisions are made at each node. Such algorithms cannot guarantee to return the globally optimal decision tree.
- Decision-tree learners create biased trees if some classes dominate. It is, therefore, recommended to balance the data set prior to fitting with the decision tree (e.g., by duplicating observations from the minority class or by randomly removing observations from the majority class).

As our data is highly imbalanced, a simple decision tree is unlikely to be the right choice for our problem.

[1] In computational complexity theory, a problem is NP-complete when:

- A non-deterministic Turing machine can solve it in polynomial-time.
- A deterministic Turing machine can solve it in large time complexity classes and can verify its solutions in polynomial time.
- It can be used to simulate any other problem with similar solvability.

XGBoost[2,3]

XGBoost stands for eXtreme Gradient Boosting.

The XGBoost library implements the gradient boosting decision tree algorithm. It is an ensembling learning method, meaning it combines several base models to produce a single optimally predictive model.

Boosting is a widely used ensemble method where new models are added to correct the errors made by existing models. Each tree learns from its predecessors and updates the residual errors. Models are added sequentially until no further improvements can be made. The base learners in boosting are weak learners in which the bias is high, and the predictive power is incrementally better than random guessing. Each of these weak learners contributes some vital information for prediction, enabling the boosting technique to produce a strong learner by effectively combining these weak learners. The final strong learner reduces both the bias and the variance.

Having a large number of trees might lead to overfitting, and, therefore, it is necessary to choose the stopping criteria for boosting carefully.

The boosting ensemble technique consists of three simple steps:

1. **Fit an initial model F_0 using the original training data (x_i,y_i) to predict the target variable y.**

$$F_0(x) = argmin_p \sum_{i=1}^{n} L\left(y_i, p(x_i)\right)$$

   where $L(y, F(x))$ is a differentiable loss function (for example $L\big(y, F(x)\big) = \|F(x) - y\|$ or $L\big(y, F(x)\big) = \|F(x) - y\|^2$) and p is a decision tree that maps the describing variables x_i to the target variable y_i.The sum is taken over all n observations. Gradient descent is used to compute the minimisation problem. This model will result in a residual $r_0 = y - F_0$:

2. **For the number of iterations m = 1,…,M:**
   - Compute the residuals r_{m-1}.
   - Fit a model f_m to the residuals r_{m-1} of the previous model.
   - Update the model: $F_m(x) = F_{m-1}(x) + \eta * f_m(x)$, where η is a given learning rate.
3. **Output $F_M(x)$**

This approach is called gradient boosting because it uses a gradient descent algorithm to minimise the loss when adding new models.

We used the function xgb.train of the XGBoost package in R:

```
xgboost_model <- xgb.train(
        params = xgb_params_typ2,
        data = training_data,
        nrounds = n,
        watchlist = watchlist_typ2,
        eta = eta_param,
        max.depth = md)
```

with the following parameters:

- params: Here you can set parameters like the number of classification classes.
- data: Training data.
- nrounds [default: 100]: Controls the maximum number of iterations.
- watchlist: Defines the output of the function.

[2] XGBoost Algorithm: XGBoost in Machine Learning. See analyticsvidhya.com.

[3] XGBoost in R : A Complete Tutorial Using XGBoost in R. See analyticsvidhya.com.

- eta [default: 0.3, range: (0,1)]: Controls the learning rate, i.e., the rate at which our model learns patterns in data. After every round, it shrinks the feature weights to reach the best optimum. Lower eta leads to slower computation (in terms of performance).
- max_depth [default: 6, range: (0, inf)]: Controls the depth of the tree. The larger the depth, the more complex the model; with higher chances of overfitting. There is no standard value for max_depth. Larger data sets require deep trees to learn the rules from data.

For the prediction of the test data the function predict can be used. It returns the probabilities of all test data for each class.

Some advantages of XGBoost are:

- High performance
- High accuracy
- Easy to use
- Can handle unbalanced data sets (where the target variable has more observations in one specific class than in the other)
- Handles large data sets well

The disadvantages of XGBoost include:

- Not as easy to interpret as a simple decision tree

Overfitting is possible when parameters are not tuned[4] properly.

There are different possibilities to measure the quality of the fitted model. We will have a look at the receiver operating characteristic (ROC) curve and the area under the curve (AUC) as described below to determine the quality of the fit.

ROC curve

An ROC curve is a performance measurement for classification problems. It is a graphical plot that illustrates the diagnostic ability of a binary classifier system, as its discrimination threshold is varied.

The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings, with:

$$TPR = \frac{TP}{TP + FN}$$

and

$$FPR = \frac{FP}{FP + TN}$$

where:

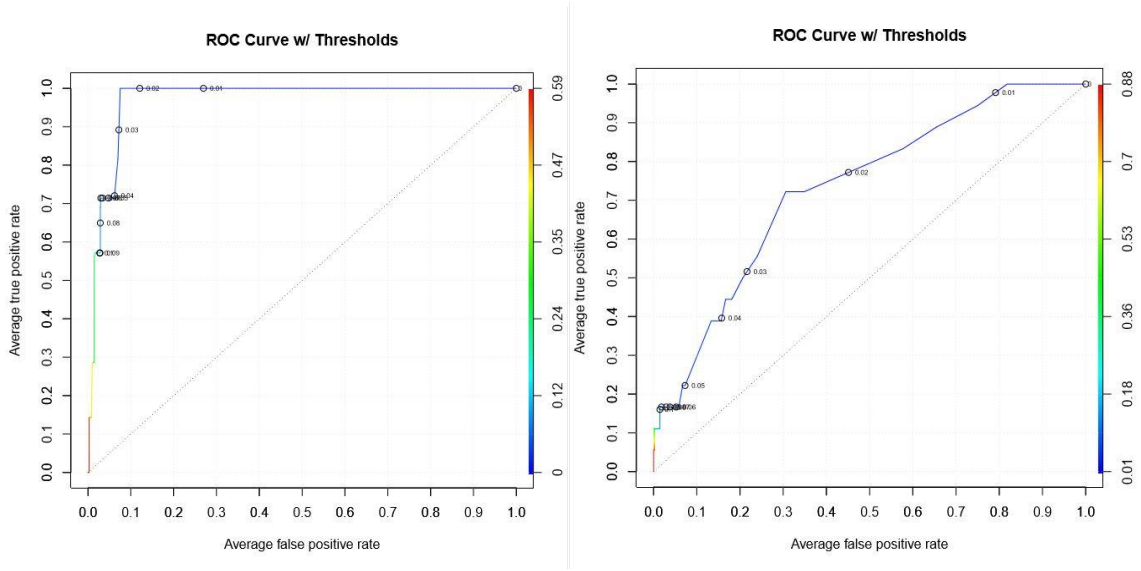
- TP: True Positives are all correctly classified as diabetics
- FP: False Positives are all incorrectly classified as diabetics
- TN: True Negatives are all correctly classified as nondiabetics
- FN: False Negatives are all incorrectly classified as nondiabetics

Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives.

In the following example, there are two different ROC curves. It is clear that the performance of the model with the left ROC curve is much better than the one on the right, as the curve is much closer to the upper-left corner.

[4] Max_depth and eta shouldn't be assumed too large in order to avoid overfitting. The model is probably overfitted if it fits the training data well but makes a poor prediction for the test data.

FIGURE 3: ROC CURVES WITH THRESHOLDS



AUC

AUC stands for "Area under the ROC curve." It measures the entire area underneath the ROC curve from (0,0) to (1,1).

The perfect model leads to $AUC = 1$ and it will have an ROC curve that passes through the upper-left corner (100% true positive rate and 0% false positive rate; in the left example of Figure 3 the AUC is nearly 1).

Results and quality of the result

There are 1,135 identified diabetics from 21,101 private-health insureds in the data set. The data set is randomly divided into training data (three-quarters of the total data) and test data (one-quarter of the total data); 861 identified diabetics are contained in the training data and 274 in the test data.

We use the XGBoost algorithm to create a model for the prediction of diabetics from the training data, depending on the three-year diagnosis histories. The model gives information about what former diagnoses are most important for a diabetes diagnosis. In the following importance matrix in the table in Figure 4, the 15 most relevant drivers for our model are shown decreasing in relevance:

FIGURE 4: IMPORTANCE MATRIX DIAGNOSES FOR TARIFF, CATEGORY 2

Position	Feature	Diagnosis.explanation
1	age_diab_2018	Age at initial diagnosis or at the start of projection
2	I48	Cardiac disease
3	E65	Consequences of obesity
4	G35	multiple sclerosis
5	I10	hypertonia
6	R70	Abnormal unspecified blood chemistry
7	H49	Strabismus paralyticus
8	K00	Unspecified disease of the teeth
9	B35	Unspecified mycosis
10	K70	Liver diseases
11	current_tariff	Tariff
12	M05	Arthropathies
13	E70	Nutritional and metabolic disorders
14	K55	vascular diseases of the intestine
15	E50	Vitamin A deficiency

It is commonly accepted that diabetes mainly occurs in mid or older ages, so it is unsurprising that age is identified in the model as the main driver for diabetes. Furthermore, there are several different main diseases which are recognised as significantly correlated to diabetes.

For example, cardiac diseases can be a result of a diabetes diagnosis but can also possibly be detected before the diabetes diagnosis.

The diagnoses E65 and E70 relate to obesity and correlate to type 2 diabetes.

Studies (5) (6) show that Vitamin A deficiency (E50) affects insulin production and is therefore directly related to diabetes.

The selected tariff gives information about the socioeconomic status of the insured person, which is correlated with health outcomes.

Some of the other listed diagnoses are known comorbidities of diabetes. They are medically proven with a correlation (like B355 or M056). Others have a high frequency or are closely related to age (like G35, H49, K00 or K55) but with no confirmed significant correlation to diabetes.

The XGBoost algorithm delivers a probability of future occurrence of diabetes for each person in the test data set.

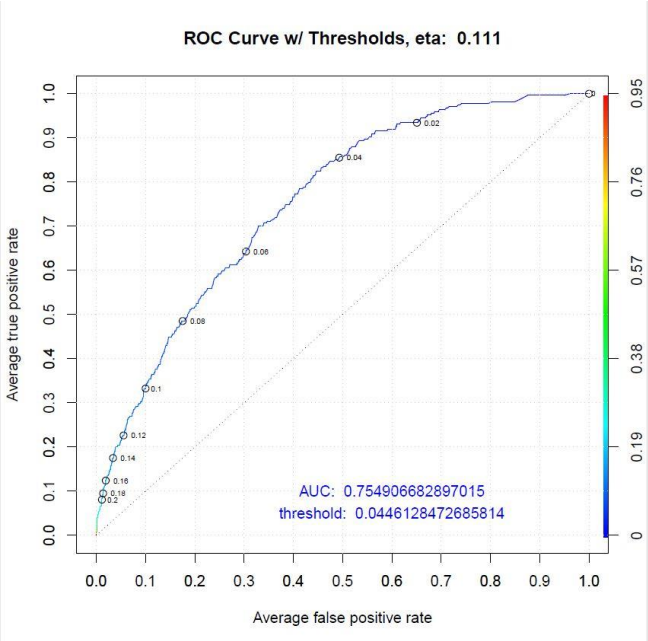
The sum of the individual probabilities reflects the expected total number of diabetics. The model estimates an expected value of 268.5 new diabetes diagnoses in the test data set, which actually contained 274 new diabetes diagnoses.

The quality of the estimation can be checked by using the ROC curve (as described above). The curve provides information on how many contracts were estimated as diabetics correctly (path of the curve in the upper-left corner) or incorrectly (path of the curve in the lower-right corner).

The ROC curve for the data set can be seen in Figure 5. In this case the path is slightly curved to the left top. This is not a perfect shape, but better than the diagonal, which only represents random guessing. The algorithm was calibrated on the data by observing the AUC. The parameter combination with the highest AUC has qualified as the best setting. The number of the iteration steps (nrounds) is the smallest nround in which the residual error no longer changes. The default for maximum depth is 6 and, as commented above, should not be too large to avoid overfitting. After testing the results of AUC, value 5 was chosen. There is also the risk of overfitting for large eta. All values from 0.001 up to the default 0.3 were tested. The best value for AUC leads to eta at 0.111.

The ROC curve can be represented as a *confusion matrix*. This shows the true positive rate (TPR) and true false rate (TFR) in relation to the threshold.

FIGURE 5: ROC CURVE FOR TARIFF, CATEGORY 2



[5] A poor diabetes control via medication promotes pathological skin changes. (8)

[6] Researchers at the University of Erlangen have identified diabetes as a risk factor for osteoarthritis, regardless of age and BMI. (9) (10)

FIGURE 6: GENERAL CONFUSION MATRIX

actual	Predicted		
	not diabetes	***diabetes***	***Row Total***
not diabetes	TN = actual_not_diabetes - FP	FP = FPR * actual_not_diabetes	actual_not_diabet
diabetes	FN = actual_diabetes - TP	TP = TPR * actual_diabetes	actual_diabetes
Column Total	predicted_not_diabetes = TN + FN	predicted_diabetes = FP + TP	test portfolio

Depending on the threshold values, the values shown in the table in Figure 7 result for the data set.

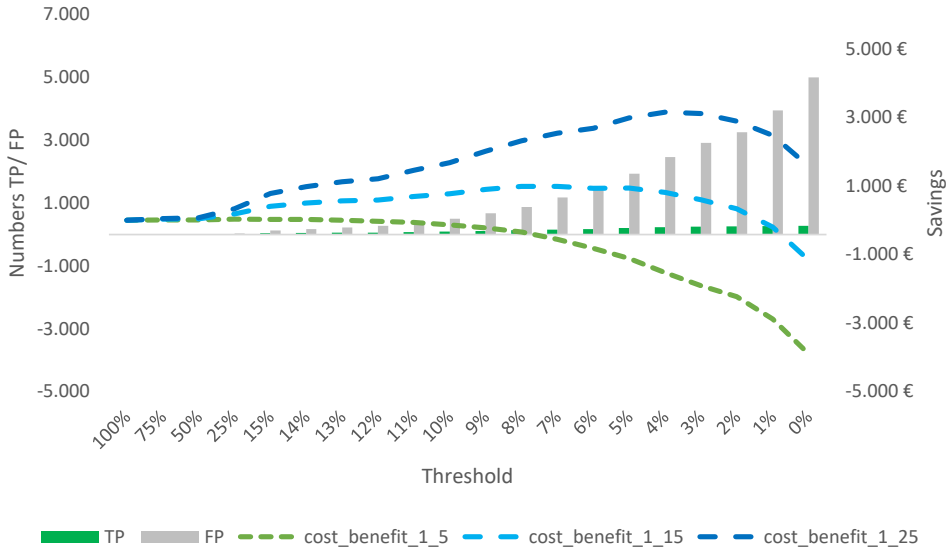
FIGURE 7: HIT RATES OF THE DATA SET

Threshold	TP	FP	TN	FN	TPR	FPR
100 %	0	0	5.001	274	0%	0%
75%	2	0	5.001	272	0%	1%
50%	3	5	4.996	271	0%	1%
25 %	15	28	4.973	259	1%	5%
15 %	38	129	4.872	236	3%	14%
10 %	91	501	4.500	183	10%	33%
5%	205	1.934	3.067	69	39%	75%
4 %	234	2.466	2.535	40	49%	85%
3 %	251	2.914	2.087	23	58%	92%
2 %	256	3.254	1.747	18	65%	93%
1 %	268	3.953	1.048	6	79%	98%
0 %	274	5.001	0	0	100%	100%

There are 274 diabetics in the test data set. In Figure 7 you can see that the higher the TP, the larger the absolute number of FPs. In the ROC curve, the perfect threshold is identified either as the value for which the ROC curve has the greatest distance to the diagonal or the smallest distance to the upper-left corner. This leads to a threshold of 4.461% and 5.662%, respectively. If we choose the optimal threshold then we get nearly 2,000 FP. This may reflect an incomplete diagnosis history because invoices are often not sent to the insurance company, due to falling below the deductible. It could also suggest that a potential diabetes diagnosis will be made in the future.

From a practical usage perspective, the threshold can be chosen based on a cost-benefit calculation. The benefit is the "saved" claims cost and the cost is the expenses for the preventive measures. If, for example, the savings are 100 times higher than the costs for the measures, then the measure can be offered to the whole portfolio, because the savings for the real diabetics are high enough that the prevention costs are not relevant. But in reality the cost-benefit ratio will be lower than 1:100, so the optimal number of attendees for prevention has to be calculated. If we assume for illustration that the savings are 5 times, 15 times and 25 times higher than the expenses for each preventive measure, respectively, then the analysis could look like what is shown in Figure 8.

FIGURE 8: COST-BENEFIT-ANALYSIS



These examples show that the higher the cost-benefit ratio, the better the prediction of the future diabetes diagnoses has to be. If the benefit related to the cost is small, then one has to reduce the costs and that means the number of FPs has to be low. If the benefit increases relative to the cost, the proportion between TPs and FPs has to become more optimal. The number of TPs must increase. Depending on the relation between the benefit for a TP and the unnecessary cost for a FP, the threshold can be chosen.

Conclusion

A good prediction of relevant diseases provides the companies with options to manage their portfolios. Chronic diseases play a crucial role in the development of future claims cost, which is why, by means of an example, we focus on diabetes type 2 in our analysis.

The analysis shows that machine learning algorithms are suitable to estimate the future development of specific chronic diseases. The main focus is on the preparation of the input data, which is important for an appropriate model. The decision-tree algorithm gives the opportunity for an explanation of the results and the importance matrix enables an insight on the drivers of a diabetes diagnosis. This is very valuable and can help in the management of future claims cost. But there are restrictions from the Federal Data Protection Act (DSGVO) to consider, because the legislation regulates data protection in Germany and gives instructions for the collection and processing of personal data. Therefore, German PHI companies are not allowed to use this analysis to offer a specific measure to a single member, though the companies could use the analysis to work out which policyholders are affected by a chronic disease and why, which may allow them to better manage sub-portfolios (e.g., with specific existing illnesses).

The algorithm described in this paper delivers an expected number of diabetics in the next year, which allows a good estimate of the upcoming claims cost. There is also the possibility of assessing preventive measure in terms of a cost-benefit approach. Last but not least, the model is very flexible in terms of the descriptive and target variables, so that there are applications for other chronic diseases or for using a prescription drugs history instead of the diagnosis history.

References

1. Javapoint.com. Decision Tree Classification Algorithm. Retrieved 12 March 2021 from https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm.
2. DAV, German Association of Actuaries (2018). Auswirkungen demografischer Effekte auf die Krankenversicherung (Ergebnispaper).
3. Health and Science (23 April 2018). Can a vitamin A deficiency promote diabetes? Retrieved 12 March 2021 from https://www.healthandscience.eu/index.php?option=com_content&view=article&id=1494:kann-ein-vitamin-a-mangel-diabetes-beguenstigen&catid=20&lang=de&Itemid=316.
4. Ibid..
5. A Deficiency Causes Hyperglycemia and Loss of Pancreatic β-Cell Mass. Retrieved 16 April 2021 from https://www.jbc.org/article/S0021-9258(20)57806-X/fulltext
6. Deutsche Diabetes Hilfe. Skin Problems and Diabetes. Retrieved 12 March 2021 from https://www.diabetesde.org/ueber_diabetes/begleiterkrankungen_bei_diabetes/hautbeschwerden.
7. Berufsverband für Orthopädie und Unfallchirurgie. Why do diabetics suffer from osteoarthritis so often? Retrieved 12 March 2021 from https://www.bvou.net/warum-leiden-diabetiker-so-oft-an-arthrose/.
8. XGBoost in R : A Complete Tutorial Using XGBoost in R (analyticsvidhya.com).
9. XGBoost Algorithm: XGBoost in Machine Learning (analyticsvidhya.com).
10. ÄrzteZeitung (14 August 2013). Type 2 diabetes increases the risk of osteoarthritis. Retrieved 12 March 2021 from https://www.aerztezeitung.de/Medizin/Typ-2-Diabetes-erhoeht-Risiko-fuer-Arthrose-282599.html.
11. Dejure.org. Insurance Supervision Act, Section 155: Changes to Premiums. Retrieved 12 March 2021 from https://dejure.org/gesetze/VAG/155.html.
12. Dejure.org. Insurance Contract Act, Section 203: Adjustment of Premiums and Conditions. Retrieved 12 March 2021 from https://dejure.org/gesetze/VVG/203.html.

Milliman

Milliman is among the world's largest providers of actuarial and related products and services. The firm has consulting practices in life insurance and financial services, property & casualty insurance, healthcare, and employee benefits. Founded in 1947, Milliman is an independent firm with offices in major cities around the globe.

de.milliman.com

CONTACT

Alexander Krauskopf
alexander.krauskopf@milliman.com

Katrin Höllering
katrin.hoellering@milliman.com

Alexandra Pflumm
alexandra.pflumm@milliman.com