Predicting Death Using Random Forests

Torsten Sauer* Roland Rau†

September 18, 2018

Abstract

Machine learning methods have become very popular in various scientific disciplines. Using Breimann's random forests and data from the National Health Interview Survey (NHIS) and its mortality follow-up, we wanted to know 1) Could these methods be used to predict the occurrence of death? 2) Which variables are important for these predictions?

We checked the accuracy of the forests by estimating the area under the ROC curve (AUC) for test data and showed that they perform relatively well, with an AUC from 0.83 to 0.87. To indicate the predictive power of every variable we estimated the mean decrease in accuracy (MDA). Not surprisingly "age" is by far the most predictive, followed by "mobility limitations" and "self-rated health". Typical sociodemographic mortality determinants like "sex", "education", and "income" seem to be very weak in their predictive ability in each of the six selected intervals.

*University of Rostock & Max Planck Institute for Demographic Research

†University of Rostock & Max Planck Institute for Demographic Research

1 Introduction

Machine learning methods have become very popular in various scientific disciplines. Prediction is one of the main applications of such techniques. For example to build and assess financial risk scores [1], to develop algorithms, which detect specific elements in genes [2], to classify patient subgroups to optimize individual treatment [3] or to predict survival of ill patients to recognise the necessity of early palliative care [4]. All of those examples need large datasets, because the accuracy of the resulting model depends on sample size as well as on the so-called feature space, i.e. the amount of predictors/variables.
Using Breiman's random forest approach [5] and data from the National Health Interview Survey (NHIS) [6], we wanted to know if these methods and data could be used to predict mortality. Therefore we asked three questions.
1) How precise can we predict a death within a given year after the interview?
2) Which variables are important for these predictions?
3) Are there variations in the predictability of these variables depending on the interval within death is predicted?

2 Data

We used nine years of NHIS data from 1997 to 2005 with 40 variables, which also include complete mortality information from the National Death Index (NDI) until the end of 2011. This led to a minimum follow up time of six years. These data have been harmonized and obtained from IPUMS [6]. We only included persons between the age of 18 and 84, who were eligble for a mortality follow up. The final data consist of 265,678 subjects. The 40 variables include demographic information about age at interview, sex, race, marital status, education, poverty, and various information about health status, i.e. self-rated health, height, weight, problems with ADL's/IADL's, hospitalizations within the last year and information about behavioral factors such as smoking, time since last doctor visit, as well as physical activity (moderate, vigor, strong).
Using mortality information about the time of death, we created six variables, which indicate if a person died (value = 1) or died not (value = 0) within each of the subsequent six years. Thus, "died not" cases could be subjects who either survived or died before

thithe selected year. These variables were used to train the random forests to predict the occurrence and timing of death. Table 1 shows that the proportion of events from 0.65% to 1.10% in all six outcome variables is very small, thereby the dataset is highly imbalanced, with respect to the outcome variables.

Table 1: Proportion of death events in intervals after the interview

Year after Interview	**No. of deaths**	**% of deaths**
1	1732	0.65
2	2394	0.90
3	2557	0.96
4	2588	0.97
5	2777	1.05
6	2926	1.10

3 Methods

We estimated six random forests [5] to predict death within each year. The advantages of random forests are: 1) they are one of the most accurate learning algorithms, 2) they can deal with imbalanced data and 3) they provide estimates of predictive power of the used variables [5, 7].

To evaluate the predictability of each forest for unseen data, we partitioned the dataset randomly into training data (80%) and test data (20%) before each fit. Because of known problems with imbalanced data, i.e. the proportion of the event of interest being relatively small, we used the SMOTE algorithm [8]. It balances the dataset before each fit. The purpose is to avoid bias in favor of the majority class [9].

The estimation of each forest consists of t estimations of so-called trees, whose results are averaged. Each fit of a tree is based on a bootstrap sample of the same size as the original sample and m randomly chosen predictors at every node of a tree. Due to its sampling with replacement, each bootstrap repetition leaves out some cases, they are called Out-Of-Bag (OOB) cases. These cases are used, after each tree fit to estimate the predictability of the hitherto trained forest resulting in the so-called Out-Of-Bag error. Based on this error we tried to find the most accurate random forests for our six

prediction intervals. Because the accuracy of a forest depends, besides the sample size and the amount of variables—which are fixed in our case—also on m, i.e. the number of randomly chosen predictors for each tree fit. We selected our random forests based on the Out-Of-Bag error evaluated by a tuning algorithm with 500 trees (t=*500*) to find the best m. Starting with eight variables to build the trees (m=*8*), we increased and decreased this number by a factor of 1.5 to see if the OBB error improved more than 1%. If this was the case, the algorithm kept looking in this direction to improve the forest further.

We used the receiver operating characteristic curve (ROC curve) to evaluate how predictive the fitted random forests were for unseen data. The ROC curve is a graphical tool, which illustrates the trade off between the amount of true positive (Sensitivity) and true negative (Specificity) predictions in a binary classifying problem, under varying decision thresholds [10]. The area under the ROC curve (AUC), a numerical value which sums up the information of the ROC curve, was used to compare the random forests with each other. The mean decrease in accuracy was used to determine the predictive power of the variables in all random forests.

The analysis was conducted in R version 3.4.4 [11] by using the "DMwR" [12] and "randomForest" [13] packages.

4 Preliminary Results

Area under the ROC curve (AUC)

Figure 1 shows the ROC curves using the estimated forests to predict death for the test data. The overall performance is given by the AUC, where 1.0 describes perfect performance classifying all test subjects to the correct class. An AUC of 0.5 would be an unprecise performance, which means that we are losing as much specificity as gaining sensitivity when adjusting the classification threshold. Sensitivity is defined as the proportion of positive—in our example "died"—cases, which were classified correctly and specificity describes the share of correctly detected negative—"not died"—cases. Decreasing the classification threshold, leads to more sensitivity and the model catches more true positve cases, but it also increases the risk of false positive classification, the model loses specificity. For example, in Figure 1 it is visible that most of the forests

are able to provide a sensitivity about 80% with just losing about 20% of specificity. Only looking at the AUC, it could be stated that our predictions are relativly good with a performance ranging between 0.83–0.87. The AUC decreases slightly with increasing time since interview (1–6 years).

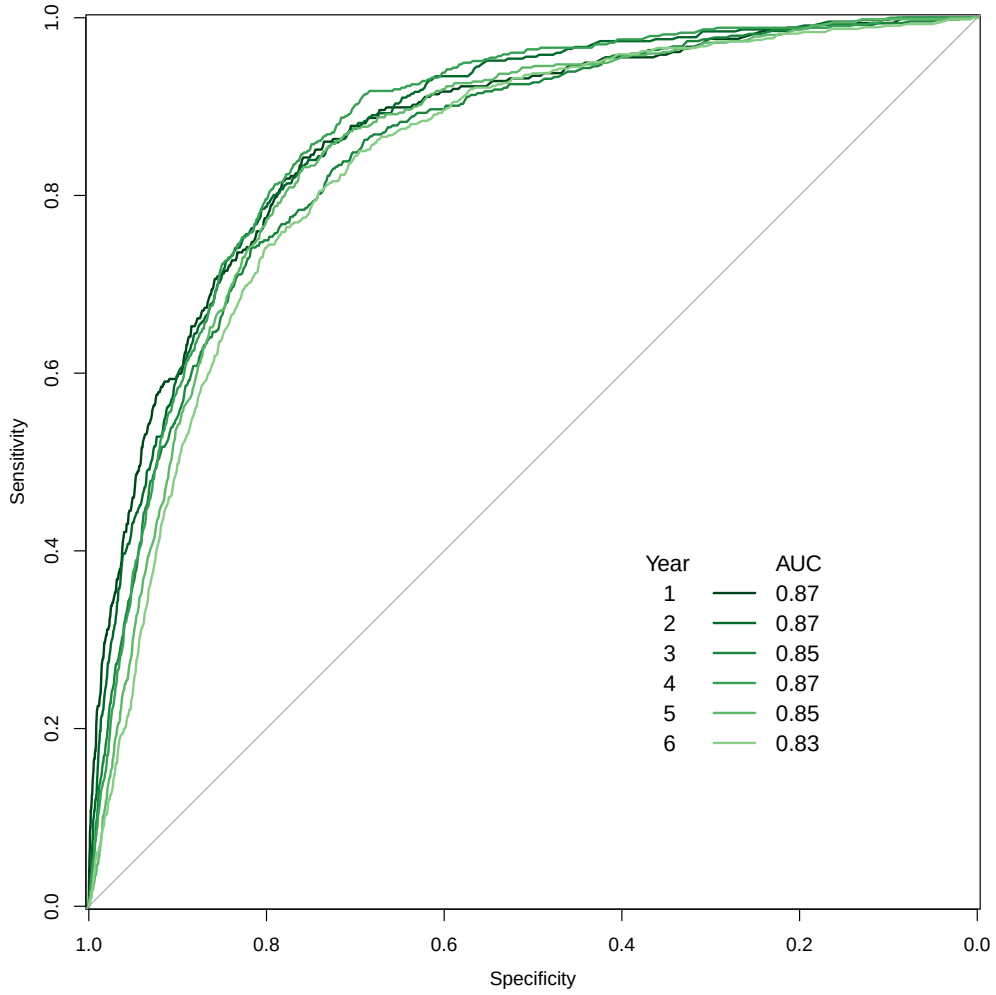

Figure 1: Areas under the ROC curves by intervals within death was predicted. Source: Own estimation and illustartion based on NHIS data (1997-2015) obtained from IPUMS [6].

Nevertheless, these results have to be handled with care. As mentioned previously, the classification variable (if someone will die in an interval or not) is highly imbalanced for our random forests. Therefore, we also show the prediction results in a so-called confusion matrix, which is a cross tabulation of the observed and the predicted cases. An example

for the prediction of death within year three after the interview is given in Table 2. Even if the specificity (80.45%) and sensitivity (74.75%) are relativly high, the precision (3.47%), i.e. the proportion of the correct classified deaths of all predicted deaths, is very low and the model estimated about 21.5 times more deaths than truly happened (10,660 to 495). Therefore we state, that only looking at the AUC as a measure of predictive accuracy in imbalanced data sets could be misleading and may result in an overestimation of the models' predictability.

Using such imbalanced datasets causes another problem. If we consider that just about 1% of our cases "die", then a simple rule of thumb, where we predict that nobody will die, would clearly outperform the random forest with its prediction error of 19.60%.

Table 2: Confusion matrix of death prediction within the third year after the interview on test data

		Prediction		
		not died	*died*	*Total*
True	*not died*	42,351	10,290	52,641
	died	125	370	495
	Total	42,467	10,660	

Test error: 19.60%; **Sensitivity:** 74.75%; **Specificity:** 80.45%; **Precision:** 3.47%

Mean decrease in accuracy (MDA)

Nevertheless, we still think that we can get interesting insights when looking at the predictability of each variable. Figure 2 illustrates which variables are important for all six predictions. The eight most predictive variables are highlighted in color. The mean decrease in accuracy could be interpreted as the average relative loss in prediction accuracy of the forest if this variable isn't considered. For example the MDA of "age at interview" in the prediction of death within the first year is 0.046, which means that we lose about 4.6% of correctly classified cases if "age at interview" was left out.

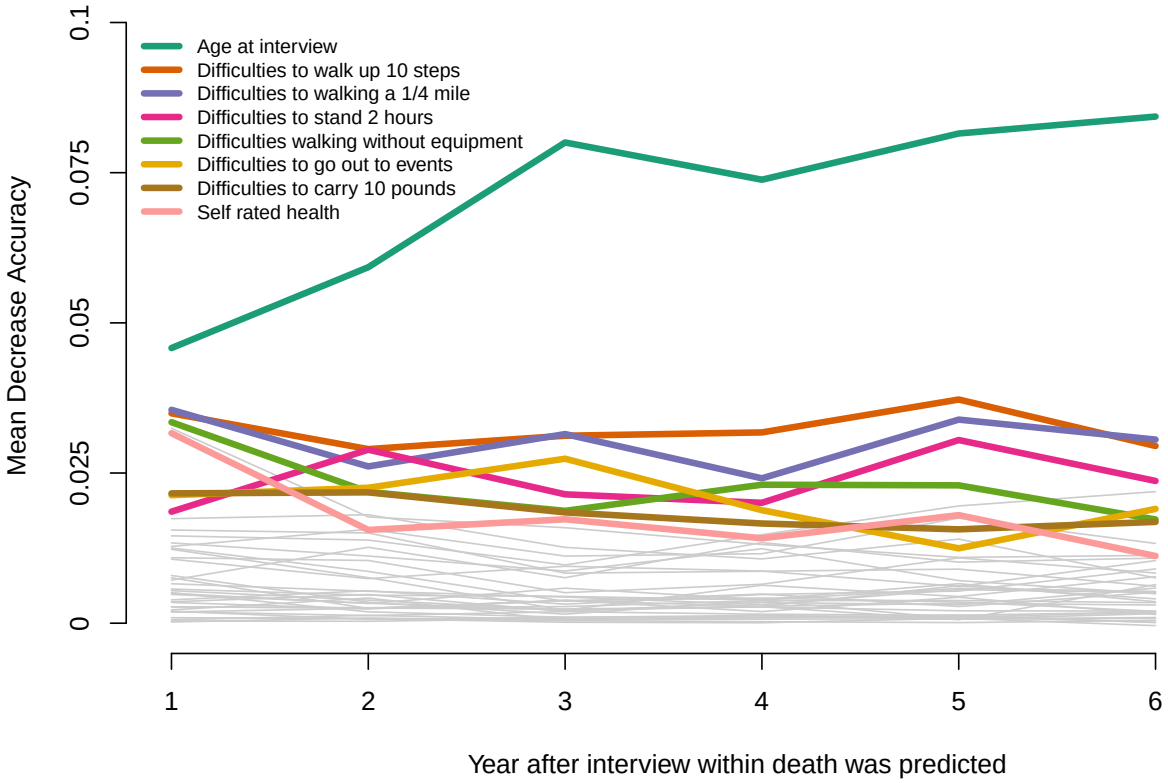

Figure 2: Mean decrease in accuracy (MDA) for all forests and all variables. The eight most predictive variables over all forest are highlighted in colors. Source: Own estimation and illustartion based on NHIS data (1997-2015) obtained from IPUMS [6].

It is not surprising that "age at interview" is by far the most important variable in all predictions and that its power is rising with increasing time since interview. Other predictive variables are information about mobility limitations, e.g in "climbing stairs", "walking a distance of a 1/4 mile" or "carrying a bag". Also "self-rated health" is one of the predictors with the highest power across all predictions. Like "hospitalizations", "self-rated health" seems to be most predictive for the first year and loses power with increasing time since interview (see also Table A).

The variables with the lowest predictability over all forests were "sex", "weight", "height", "poverty", "ethnicity", "health insurance coverage", "numbers of doctor visits","health compared to one year ago" and "numbers of surgeries in the last year".

While we can explain ourselves why age, mobility limitations and health are good predictors. We wonder about the weakness in predictability of typical socidemographical mortality determinants such as sex, ethnicity, education and income.

References

[1] J. Galindo and P. Tamayo, "Credit risk assessment using statistical and machine learning: Basic methodology and risk modeling applications," *Computational Economics*, vol. 15, pp. 107–143, Apr 2000.

[2] M. W. Libbrecht and W. S. Noble, "Machine learning applications in genetics and genomics," *Nature Reviews Genetics*, vol. 16, pp. 321–332, may 2015.

[3] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and Structural Biotechnology Journal*, vol. 13, pp. 8–17, 2015.

[4] A. Avati, K. Jung, S. Harman, L. Downing, A. Ng, and N. H. Shah, "Improving palliative care with deep learning," *ArXiv*, 2017. URL:http://arxiv.org/abs/1711.06402v1.

[5] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, Oct 2001.

[6] L. A. Blewett, J. A. R. Drew, R. Griffin, M. L. King, and K. C. Williams, "IPUMS Health Surveys: National Health Interview Survey, Version 6.3," 2018.

[7] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer New York, 2009.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, jun 2002.

[9] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Editorial: Special issue on learning from imbalanced data sets," *ACM SIGKDD Explorations Newsletter*, vol. 6, p. 1, jun 2004.

[10] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. Springer New York, 2013.

[11] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.

[12] L. Torgo, *Data Mining with R, learning with case studies*. Chapman and Hall/CRC, 2010.

[13] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

A Appendix

Table A: Mean decrease in accuracy (MDA) in percent for prediction of death within year 1, 3 and 6 order by their power of prediction

	Year 1		Year 3		Year 6	
Nr.	**Variable**	**MDA in %**	**Variable**	**MDA in %**	**Variable**	**MDA in %**
1	Age at interview	4.58	Age at interview	8.00	Age at interview	8.43
2	Difficulties walking 1/4 mile	3.55	Difficulties walking 1/4 mile	3.15	Difficulties walking 1/4 mile	3.06
3	Difficulties walking 10 steps	3.49	Difficulties walking 10 steps	3.12	Difficulties walking 10 steps	2.95
4	Difficulties walking without equipment	3.35	Difficulties to go to events	2.74	Difficulties to stand for 2h	2.37
5	Hospitalizations	3.24	Difficulties to stand for 2h	2.15	Difficulties to stoop/bend/kneel	2.19
6	Self rated health	3.16	Difficulties walking without equipment	1.87	Difficulties to go to events	1.90
7	Difficulties to carry 10 pounds	2.16	Difficulties to carry 10 pound	1.84	Difficulties walking without equipment	1.73
8	Difficulties to go to events	2.13	Self rated health	1.73	Difficulties to carry 10 pounds	1.69
9	Difficulties to stand for 2h	1.86	Hospitalizations	1.59	Difficulties to push large objects	1.33
10	Difficulties social activities	1.74	Difficulties social activities	1.26	Hospitalizations	1.13
11	Need of special equipment	1.55	Difficulties to push large objects	1.12	Self rated health	1.12
12	Difficulties to stoop/bend/kneel	1.45	Difficulties to stoop/bend/kneel	0.97	Marital status	1.08