Predicting hospital associated disability from imbalanced data using supervised learning

Saarela, Mirka; Ryynänen, Olli-Pekka; Äyrämö, Sami

Year: 2019

Version: Accepted version (Final draft)

Copyright: © 2018 Elsevier B.V.

Rights: CC BY-NC-ND 4.0

Rights url: https://creativecommons.org/licenses/by-nc-nd/4.0/

Please cite the original version:

Predicting Hospital Associated Disability from Imbalanced Data Using Supervised Learning

Mirka Saarela\textsuperscript{a*}, Olli-Pekka Ryynänen\textsuperscript{b,c}, Sami Åyrätno\textsuperscript{a}

\textsuperscript{a}University of Jyväskylä, Faculty of Information Technology, P.O. Box 35, FI-40014 University of Jyväskylä, Finland
\textsuperscript{b}Department of Public Health and Clinical Nutrition, University of Eastern Finland, Kuopio, Finland
\textsuperscript{c}General Practice Unit, Kuopio University Hospital, Primary Health Care, Kuopio, Finland

Abstract

Hospitalization of elderly patients can lead to serious adverse effects on their functional capability. Identifying the underlying factors leading to such adverse effects is an active area of medical research. The purpose of the current paper is to show the potential of artificial intelligence in the form of machine learning to complement the existing medical research. This is accomplished by studying the outcome of hospitalization of elderly patients as a supervised learning task. A rich set of features characterizing the medical and social situation of elderly patients is leveraged and using confusion matrices, association rule mining, and two different classes of supervised learning algorithms, it is shown that the need for help and supervision are the most important features predicting whether these patients will return home after hospitalization. Such findings can help to improve hospitalization and rehabilitation of elderly patients.

Keywords: Hospital Associated Disability, Machine Learning, Random Forest

1. Introduction

Basic activities of daily living (short ADLs) are fundamental to ensure that older people are able to live independently without care. These ADLs include activities such as bathing, dressing, using a toilet, and eating. Hospitalization of previously independent older patients due to an acute medical illness often leads to a situation in which the patients find themselves unable to perform one of these activities independently anymore [1]. Consequently, they cannot return to their old lives but are dependent on help also after their stay at the hospital. This undesirable side-effect is called hospital associated disability (HAD).

HAD is associated with the advent of serious adverse events, such as mortality, morbidity, institutionalization and re-hospitalizations, affecting both the individuals (reduced quality of life or even death) and the society (e.g., higher use of caregivers and other resources, greater health care expenditure). It can be actuacted although the illness that necessitated the hospitalization was successfully treated [2, 3, 4] and although the patients admission diagnosis was not related to a decline in ADLs [5]. Saltvedt et al. [6] showed that geriatric evaluation and management units where older patients are encouraged to participate in ADLs during hospitalization significantly reduce mortality of the patients. However, according to Covinsky et al. [7], approximately one-third of elderly patients have an ADL disability that they did not have before the hospitalization. Thus, HAD is a serious problem and understanding the most important factors leading to it could help the elderly and save many resources.

The purpose of this paper is to find accurate prediction methods for the outcome of hospitalization of elderly patients and to recognize the strongest factors predicting HAD. Earlier studies have pointed out that high age, comorbid disease, depression, cognitive impairment, limited social support and physical frailty can cause HAD [1, 2, 7, 8]. Table 1 provides an overview of previous studies concerned with predicting HAD. It summarizes for all of these studies the analyzed data, used methods, and main findings. As can be seen from the table, all of the existing studies used linear techniques originating from statistics such as t-tests, $\chi^2$ tests, or logistic regression.

However, comparative research on supervised learning algorithms has shown that linear methods such as...
logistic regression are not competitive with and constantly outperformed by more complex and nonparametric methods such as random forest [9, 10, 11]. Random forest has been a preferred choice especially in many medical applications because it not only shows very good prediction performance but is also known for its ability to identify important variables [12]. This ability to identify important variables is generally a crucial property for health care applications as it increases the interpretability of the models, which helps medical decision makers to understand and use these models as clinical decision support [13, 14]. Another reason why random forest may be advantageous is that it has proved to be more robust toward imbalanced data. This is important here because—as in many other, particularly medicine related, applications—our real world data set is imbalanced and misclassifying a minority class observation is more serious than misclassifying a majority class observation.

Hence, our analysis adds to the existing studies in two respects. First, we use random forest to automatically learn a model that best fits the data. We show that this model outperforms the logistic regression model. Second, we use methodological triangulation of different analysis techniques to improve the technical soundness of the presented approaches [15]. This triangulated analysis gives more confidence on our final conclusion that previously estimated needs for help and supervision are the most important predictors for HAD.

2. Data

The data was prospectively collected from four national hospitals in Finland located in Helsinki, Joensuu, Jyväskylä and Kuopio. Only patients who arrived from a hospitalization were defined (i) as the categorical variable discharge to home, institutionalized, or dead, and (ii) simply as binary variable distinguishing discharge to home or institutionalized or dead. For three cases in our data the dependent variable was missing. Since we needed the information of hospitalization outcome in all our analysis techniques those three cases were deleted.

The patients who did not return home were designated to be the cohort of interest (i.e., the ones with HAD). Thus, the distribution of patients with HAD in the data set was very similar to the general occurrences of HAD estimated by Covinsky et al. [7]. One third (i.e., 285) of the patients in our dataset were either institutionalized or dead, exactly as the estimated occurrence of HAD (see Section 1).

2.1. Features

The collected data features 100 variables, two dependent and 98 independent features related to the patients health and social status. 92 of the 98 original features are categorical. The majority of these features are ordinal (Likert-scale) such as walking stairs that has the categories without difficulty, with help, and not at all. However, there are also some purely categorical variables in the collected data, such as gender and location. All original variables were transformed into numerical features, either real-valued or binary ones. Moreover, as described more detailed below, the absence of a value was encoded with a separate category for each variable because it can provide valuable information.

For some of our analysis techniques (see Section 3) we need our whole data in binary format. For this, we categorized the six non-categorical features (age, hemoglobin, white blood cell count, glucose level, sodium level, potassium level) and one-hot encoded all variables with an additional variable indicating the missing information. This led to a target data with 381 features. Removal of constant features left us with 332 binary features.

2.2. Missing data

Altogether, our dataset has less than 3% of missing data. However, 53% of the variables are incomplete. These missing values occur in 701 (84%) of the observations. That means that for only 16% of the patients we have values for all variables. Furthermore, the missing data are not evenly distributed variablewise. On the one hand, a lot of features have only a couple (less than four) or no missing values. On the other hand, a few features have many missing values. One example of the latter case is the glucose level, which has more than 50% missing values. The absence of a glucose level for a patient in our data set usually indicates that based on previous examinations the value was assumed to be within the normal range. That means that some features in our data are reported only for certain patients with certain conditions and the occurrence of a missing value is actually related to the reason why it is missing. Thus, the data are missing not at random (MNAR) [20].

To deal with the MNAR sparsity pattern we encoded for each feature that had missing values a new category
Table 1: Related work

<table>
<thead>
<tr>
<th>Authors</th>
<th>Data</th>
<th>Method</th>
<th>Main findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zisberg et al. [16]</td>
<td>684 patients aged 70 or older of two Israeli tertiary medical centers</td>
<td>correlation analysis, path analysis</td>
<td>Mobility inside the hospital, continence care, and length of stay were associated with HAD.</td>
</tr>
<tr>
<td>Lees et al. [17]</td>
<td>257 patients aged 65-80 undergoing emergency general surgery at the University of Alberta Hospital in Canada</td>
<td>logistic regression</td>
<td>Elderly patients discharged to home after hospitalization were younger, had fewer in-hospital complications and lower American Society of Anesthesiologists class.</td>
</tr>
<tr>
<td>Covinsky et al. [2]</td>
<td>2293 patients aged 70 or older from University Hospitals of Cleveland, USA in Ohio, USA</td>
<td>(\chi^2) tests and logistic regression</td>
<td>The oldest patients were the most likely to develop a HAD.</td>
</tr>
<tr>
<td>Gill et al. [1]</td>
<td>754 community-living persons aged 70 years or older who were nondisabled at baseline, in New Haven, Connecticut, USA</td>
<td>confidence intervals, Cox model</td>
<td>Hospitalization of the elderly was strongly associated with loss of an ADL, especially for frail individuals.</td>
</tr>
<tr>
<td>Carlson et al. [18]</td>
<td>122 patients aged 60 or older of an acute care geriatric inpatient unit of university hospital in Texas, USA</td>
<td>t-test, (\chi^2) tests, logistic regression</td>
<td>Poor functional homeostasis was significantly associated with HAD independently of other patient’s characteristics.</td>
</tr>
<tr>
<td>Wu et al. [19]</td>
<td>804 patients aged 80 years or older who stayed in a USA hospital (Beth Israel Hospital, Boston; MetroHealth Medical Center, Cleveland; Marshfield Clinic/St. Joseph’s Hospital, Marshfield; University of California Los Angeles Medical Center, Los Angeles) at least 48 hours</td>
<td>ordinal logistic regression models</td>
<td>The strongest independent predictor of HAD was the ADL score at baseline. For patients independent in ADLs at baseline, the presence of an orthopedic diagnosis was associated with poorer subsequent function. However, for patients with four or more ADL dependencies at baseline, orthopedic diagnoses was not independently associated with HAD.</td>
</tr>
<tr>
<td>Inouye et al. [5]</td>
<td>188 patients aged 70 or older admitted to the medicine service at Yale-New Haven hospital, USA</td>
<td>t-test, (\chi^2) tests, proportional hazards model</td>
<td>The risk of functional decline in ADLs increased linearly with the number of risk factors, suggesting that the predisposition to functional decline may result from the cumulative effects of multiple impairments.</td>
</tr>
</tbody>
</table>

indicating the missing information. This strategy was chosen for two reasons. First, it ensures that we do not need imputation, which is difficult or even impossible with a MNAR pattern. Second, missingness patterns might reveal interesting information itself. For example, as explained above, the glucose level was collected and reported mainly for patients with non-normal levels. Hence, the missingness of the patient’s glucose level is an indicator that this level was in normal range. Moreover, as illustrated in Figure 1 the distribution of discharged to home patients is different for missing values (i.e., when there are missing values, less patients are discharged to home than usually). We do not want to delete this information but uncover it in case it is related to the outcome of hospitalization. Hence, for each variable with missing values a new category indicating the availability of this features was created.

2.3. Class imbalance

Another challenge of our data set is the class imbalance of the collected data. In various application areas—especially medicine—the class of interest (e.g., patient having a disease) is underrepresented in the data while the majority of the data represent the control cases. Classifiers built on such imbalanced data often are biased toward the majority class [21]. As in many real medical data sets we have less cases of the class of interest (i.e., one third of the patients belong to the HAD class) than control cases (i.e., the class of return home patients). This means that the trivial classifier that sim-
(a) Distribution of the fracture variable.

(b) Distribution of the vomiting variable.

(c) Distribution of the need for help variable.

(d) Distribution of the pain variable.

Figure 1: Distribution of discharged to home patients for selected variables. The missing values (denoted as −1) appear more often with patients that were not discharged to home.

ply assigns every patient to the most common (returning home) class would achieve an accuracy of 66% while having 0% sensitivity.

The class imbalance also means that we cannot evaluate our classification models based on the accuracy. Thus, instead of concentrating on a high accuracy we compare our classification models based on the area under the ROC curve (AUC). The AUC is more independent of class skew because it measures the ranking abilities of the classifier. It is equivalent to the probability that the classifier will rank a randomly selected positive observation (in our case, HAD patient) higher than a randomly selected negative observation (i.e., a returning home patient in our case). Hence, an AUC value around 0.5 is no better than random guessing while an AUC of 1 presents perfect classification.

To deal with the class imbalance problem in classification different procedures have been introduced (for a recent review see Vanhoeyveld and Martens [22] and references therein). These procedures can mainly be divided into procedures that operate on the classification algorithm (cost-sensitive methods) and procedures that operate on the data (sampling methods). Cost-sensitive methods assign larger weights to observations of the minority class. As a consequence, minority class misclassification is more costly which impedes the trivial systematic classification to the majority class. Sampling methods modify the training set by sampling a smaller majority training set (undersampling) or repeating observations from the minority training set (oversampling).

A popular sampling technique of the latter case is synthetic minority oversampling (SMOTE) [23]. SMOTE creates observations that are synthetic interpolations of
the minority class. This produces clusters around each observation from the minority class instead of simply duplicating observations (such as the random oversampling method does). In a recent study by Bach et al. [24], different solutions to deal with imbalanced data were compared in the context of medicine (i.e., prediction of osteoporosis). They experimented with different classifiers and procedures to cope with their imbalanced data and showed that oversampling the minority (i.e., osteoporosis) class using SMOTE resulted in higher AUC values for all tested classifiers.

Liu et al. [25] introduced the EasyEnsemble method. The EasyEnsemble method samples from the training set as many observations from the majority class as there are observations from the minority class in the training set. This procedure is repeated until each majority class observation is sampled into one training set. Then, each balanced training set is used to train the classifier and the outputs of those learners are combined for the prediction. This method was tested on classification tasks for imbalanced data sets and found to outperform other sampling techniques by Vanhoeyveld and Martens [22].

3. Methods

Our analysis was based on a combination of different data mining and machine learning techniques [26], all with the goal to find the most important predictor for HAD from the data:

- Confusion matrix to identify the single variables that predict HAD the best
- Association rule mining to determine interesting patterns in the data that co-occur (i) with HAD patients and (ii) with the patients that return home
- Supervised learning to (i) find the model that predicts HAD the best using all information from the data including the missingness information and (ii) identify the features most useful for the supervised models utilizing feature importance measures

3.1. Direct effect of single variables using confusion matrices

First, we determined the importance of the 332 binary features originating from the one-hot-encoding described in Section 2. For each of these features, we computed the confusion matrix of this single feature with the dependent variable institutionalized or dead. We then used the formula for accuracy (i.e., the sum of true positives and true negatives divided by the number of observations) to determine the importance of the variables. All variables that had a higher accuracy than the default classifier that always predicts the most common class (i.e., discharge to home) and is right in 66% (546 out of 832) of the cases, were considered to be informative.

3.2. Association rule mining

Second, we used association rule mining to automatically detect what patterns and if-then rules could be found in the binarized data. In association rule mining, these pattern and rules are usually represented in forms of implication rules [26, 27]. If $I$ is the set of all items and $S_m$ a subset of $I$ ($S_m \subseteq I$) then a transaction $t_i \in T$, where $T$ denotes the set of all transactions, is said to contain $S_m$ if $S_m$ is a subset of $t_i$. The number of all transactions of a dataset that contain a particular itemset is its support count. It is defined as $\sigma(S_m) = |\{ t_i \mid S_m \subseteq t_i \in T\}|$, where $|\cdot|$ stands for the number of elements in a set.

An association rule is an implication expression of the form $S_m \rightarrow S_n$, where $S_m, S_n \subseteq I$ and $S_m \cap S_n = \emptyset$. The goodness of such a rule is typically assessed by its support and confidence. The support determines how often a rule is applicable to a given data set. It is defined as $s(S_m \rightarrow S_n) = \frac{\sigma(S_m \cup S_n)}{|T|}$. The confidence assesses how frequently items in $S_n$ appear in transactions that contain $S_m$ and is defined as $c(S_m \rightarrow S_n) = \frac{\sigma(S_m \cup S_n)}{\sigma(S_m)}$.

3.3. Supervised learning using logistic regression and random forest

Third, we wanted to predict HAD using supervised learning. As identified in Table 1, logistic regression seems to be still the standard technique to predict HAD. We believe that this due to two reasons. First, logistic regression is known for its straightforward interpretability [13, 14], which is a property that is (as discussed in Section 1) of great importance in medical applications. Second, logistic regression is probably the most established multivariate prediction technique for binary outcomes in statistics.

Our other supervised learning technique is random forest. Previous research has highlighted the great potential of random forest for biomedical applications [28, 29]. They are one of the best state-of-art classifiers known in machine learning and require only very little parameter tuning [30, 31]. Moreover, random forests are often preferred choices in biomedical applications because of their embedded feature importance measure which facilitates interpretability [12]. For example, as
probability of class $c_i$ defined as trees by bagging many of such trees and introducing two the overfitting problem of single fully grown decision the average of a multitude of uncorrelated trees is reduced prediction performance on the test set. (i.e., su
ting, the average of a multitude of uncorrelated trees is

$$G(X) = 1 - \sum_{i=1}^{C} P(c_i|X)^2,$$

where $C$ denotes the number of classes and $P(c_i|X)$ the probability of class $c_i$ in $X$. A Gini index value of zero is gained if all observations in a data partition belong to the same class (the probability of that class is 1 while the probability of all other classes is 0), while larger values indicate less pure class divisions. Thus, at each node the selected split is the one with the lowest Gini index.

Decision trees handle high-dimensional data well and are—although nonlinear—highly amenable to human model interpretation as they provide understandable rules on the splitting attributes. A further strong point of decision trees is their ability to handle mixed data types such as categorical and continuous features. However, if fully grown, they are prone to overfitting (i.e., suffer from high variance) and thus often lead to reduced prediction performance on the test set.

While a single decision tree is highly prone to overfitting, the average of a multitude of uncorrelated trees is not. Consequently, a random forest classifier solves the overfitting problem of single fully grown decision trees by bagging many of such trees and introducing two sources of randomness: (i) randomness in the data and (ii) randomness in the features. Bagging is a process in which each decision tree is constructed from a bootstrap sample drawn from the original data set and the prediction of a new observation is made by taking the mode (average in case of regression) of all trees. Thus, randomness in the data is the result of bagging since the bootstrap samples ensure that the trees in the forest are built on different training sets. Randomness in the features is the result of considering for each split of a decision tree only a small number of features $q$, with $q$ being a user defined parameter and typically much smaller than the total number of features $p$ (default $p = \sqrt{q}$).

If many trees were trained on the same data and if all features were considered in each split, the trees of such a forest were strongly correlated or even duplicates of each other. However, the trees of a random forest are not correlated since each tree in the random forest model is grown on a different set of data and considers only a subset of the features as splitting attributes on each node. Hence, the random forest leads to improved performance because it decreases the variance (overfitting) of the model without increasing the bias. As such, they are frequently the winning classifiers in machine learning competitions.

4. Results

Below we report the results of our different analysis techniques outlined in Section 3. As explained we were not only interested in a prediction model with high performance but also in finding most explanatory variables expressing what characteristics of elderly patients are associated with so permanent adverse effects on functionality that they cannot be discharged anymore, but need long-term care or even die after being hospitalized for reason of acute illness. Thus, we also discuss for each technique the most important predictors.

All experiments were implemented in Python 3.6.2 (using scikit- and imbalanced-learn packages) and Matlab R2016b for association rule mining.

4.1. Direct effect of single variables using confusion matrices

All informative predictor variables (i.e., features that provide a clearer distribution than the default) are listed in Table 2. As can be seen from the table, 56 out of the 332 features were identified to be more informative than the default predictor and the need for supervision was the most important variable with a prediction accuracy of 67% (560 out of 832 cases).
Table 2: List of one-hot encoded features that were more accurate in prediction than the default classifier that always predicts the most common class and is correct in 546 out of 832 (66%) of the cases.

<table>
<thead>
<tr>
<th>variable name</th>
<th>correctly classified</th>
<th>variable name</th>
<th>correctly classified</th>
</tr>
</thead>
<tbody>
<tr>
<td>need_for_supervision_2.0</td>
<td>560</td>
<td>pain_nan</td>
<td>559</td>
</tr>
<tr>
<td>remembering_new_things_3.0</td>
<td>559</td>
<td>need_for_help_nan</td>
<td>558</td>
</tr>
<tr>
<td>confusion_3.0</td>
<td>556</td>
<td>eating_2.0</td>
<td>555</td>
</tr>
<tr>
<td>need_for_help_how_often_2.0</td>
<td>555</td>
<td>mobility_aids_nan</td>
<td>554</td>
</tr>
<tr>
<td>falling_3.0</td>
<td>551</td>
<td>shakiness_3.0</td>
<td>551</td>
</tr>
<tr>
<td>confusion_4.0</td>
<td>551</td>
<td>mobility_aids_4.0</td>
<td>551</td>
</tr>
<tr>
<td>living_before_admission_nan</td>
<td>550</td>
<td>alzheimer_2.0</td>
<td>551</td>
</tr>
<tr>
<td>depression_2.0</td>
<td>550</td>
<td>mobility_aids_5.0</td>
<td>550</td>
</tr>
<tr>
<td>confusion_2.0</td>
<td>550</td>
<td>mental_disorder_2.0</td>
<td>550</td>
</tr>
<tr>
<td>psychosis_2.0</td>
<td>549</td>
<td>need_for_help_4.0</td>
<td>548</td>
</tr>
<tr>
<td>underweight_2.0</td>
<td>548</td>
<td>arthralgia_nan</td>
<td>548</td>
</tr>
<tr>
<td>heartburn_nan</td>
<td>548</td>
<td>stroke_3.0</td>
<td>548</td>
</tr>
<tr>
<td>walking_stairs_nan</td>
<td>548</td>
<td>need_for_supervision_nan</td>
<td>548</td>
</tr>
<tr>
<td>alcohol_2.0</td>
<td>548</td>
<td>numbing_of_bag_nan</td>
<td>548</td>
</tr>
<tr>
<td>back_pain_nan</td>
<td>547</td>
<td>oedema_nan</td>
<td>547</td>
</tr>
<tr>
<td>dyspnoea_in_strain_nan</td>
<td>547</td>
<td>dyspnoea_in_rest_nan</td>
<td>547</td>
</tr>
<tr>
<td>excema_nan</td>
<td>547</td>
<td>constipation_nan</td>
<td>547</td>
</tr>
<tr>
<td>dizziness_nan</td>
<td>547</td>
<td>falling_nan</td>
<td>547</td>
</tr>
<tr>
<td>shakiness_nan</td>
<td>547</td>
<td>stroke_nan</td>
<td>547</td>
</tr>
<tr>
<td>speak_difficulty_3.0</td>
<td>547</td>
<td>speak_difficulty_nan</td>
<td>547</td>
</tr>
<tr>
<td>walking_room_3.0</td>
<td>547</td>
<td>walking_room_nan</td>
<td>547</td>
</tr>
<tr>
<td>walking_outside_4.0</td>
<td>547</td>
<td>walking_outside_nan</td>
<td>547</td>
</tr>
<tr>
<td>walking_and_carrying_bag_nan</td>
<td>547</td>
<td>using_toilet_nan</td>
<td>547</td>
</tr>
<tr>
<td>finger_dexterity_nan</td>
<td>547</td>
<td>washing_nan</td>
<td>547</td>
</tr>
<tr>
<td>cutting_toe_nails_nan</td>
<td>547</td>
<td>eating_nan</td>
<td>547</td>
</tr>
<tr>
<td>mobility_aids_2.3</td>
<td>547</td>
<td>need_for_supervision_1.0</td>
<td>547</td>
</tr>
<tr>
<td>multimorbidity_binary_2.0</td>
<td>547</td>
<td>delusions_2.0</td>
<td>547</td>
</tr>
<tr>
<td>gastrointestinal_symptoms_nan</td>
<td>547</td>
<td>loss_of_balance_nan</td>
<td>547</td>
</tr>
</tbody>
</table>

4.2. Association rules

Our goal with the second technique was to find rules of strongly associated features in our data that indicate that a patient is in risk to be institutionalized or die. Since only one third of the observations in our data actually contain patients that died or were institutionalized, we started with a relatively small value for the support but the highest value for the confidence to achieve reliable and accurate rules.

Table 3 shows the result when setting the support to 0.6 and searching for rules with HAD on the right hand side of the rule with as high confidence as possible. The first rule that we obtained with this strategy had a confidence of 100%. All patients that needed help one to five times a week and for which pain was missing, were institutionalized or died (see rule number 1 in Table 3). Similarly, all underweight patients for whom the need for help was missing were institutionalized or died (rule 2 in Table 3).

The following rules did not have full confidence but a higher support. More than 64% of those patients with Parkinson who needed help one to five times a week (1.1% of the data) were institutionalized or died (rule 3). As already detected with confusion matrices, the need for help appeared again to be a good indicator for HAD. More than half of the patients who needed help one to five times a week (21.3% of all patients in the data) were institutionalized or died (rule 8). This percentage is even higher if the patient lived alone before admission to the hospital (rule 7). The alone (living) situation in combination with the need for help is also reflected by rule 4 and 5. If the patient is single or a widow(er) and he or she needs help or suffers from confusion, HAD is more likely. If we search for interesting rules with home at the right hand side, we obtain rule 9 stating that if a patient does not need help (note both binarized categories of help demand variables measure the same), he or she will probably (confidence more than 93%) return home.

In summary, association rule mining reflected the importance of the variables that measure the amount of help the elderly patients needed already before hospitalization. If a patient did not need help or supervision, he or she will probably return home. If a patient needed some help, it is likely that he or she will be institutionalized or die after hospitalization. This is especially true if the patient is alone (i.e., lives alone and/or is single or widowed).
4.3. Supervised learning

We finalized our triangulated analysis by predicting HAD using supervised learning. As pointed out in the introduction, we were especially interested in the performance of the random forest classifier in comparison to the logistic regression classifier because this appears to be the main (exclusive) classification model in medical research to predict HAD (see Table 1).

In order to assess the performance of the different models the data set was split into training (70%) and test set (30%). Grid-search hyperparameter optimization with 10-fold cross-validation was used to evaluate the individual models on the training set. The final model using the best performing combination of hyper-parameters was then refit on the full training data set and the performance was evaluated on the independent test set. The test set was untouched during the entire training and model selection process and only used for the final model evaluation. Therefore, it was ensured that no information of the test set was revealed during model training and parameter optimization.

As discussed above our data is challenging because of the missing data and the class imbalance. To find what works best for our classification problem and our algorithms we experimented with all strategies discussed in Section 2: the cost-sensitive procedure (using balanced class weights in the classification algorithm), different sampling of the training data (i.e., random oversampling, SMOTE, random undersampling, and EasyEnsemble), and none for comparison.

A random forest with 1000 estimators\(^1\) and Gini impurity (see 1) as splitting criterion was used. The maximum number of features was grid searched. For logistic regression the best penalty and regularization parameter were grid searched. Moreover, we searched for the best dimension reduction (principal component analysis, non-negative matrix factorization, or none) and best scaling (standardization, min-max-scaling, or none).

Table 4 shows the experimental results. Pipelines were used in combination with the grid search to chain the preprocessing (scaling and dimension reduction) steps together with the classifiers and to ensure that no information from the test set was leaked into the training set. That means that we encapsulated all steps in a single estimator and automatically tried all possible combinations of the specified preprocessings and classifier parameters without touching the test set. Thus, the final parameters and preprocessing reported in the table and applied to the test set correspond to the ones returned by scikit-learn’s pipeline and grid search (i.e., best_params from the GridSearchCV).

The different preprocessings tested for our models gave consistently the same result. The random forest models performed the best with no preprocessing (no scaling and no dimension reduction) and the logistic regression with standard scaling and no dimension reduction (see column best preprocessing in Table 4). Similarly, there was only little difference between the tested configurations regarding the grid searched values of hyperparameters (see column best parameters in Table 4). For example, for all random forest models we obtained the best results when only a small number of maximum features were considered at each split (i.e., the best value for \(q\) in our random forest models always was either 3 or 4, see rows 1–6 in Table 4). We further observed that our logistic regression models always performed better with L2 than with L1 regularization (see rows 7–12 in Table 4). Although L1 regularization could probably lead to better model interpretability by zeroing weights (pruning variables) with higher probability than L2 regularization, the results indicate that the

---

\(^1\)In the beginning, the number of trees was also grid searched. However, more trees were always better and we observed that after 1000 trees there was no gain in performance anymore but a longer computation time.
latter one, which assumes zero-mean Gaussian prior distribution over the weights, produces models with greater predictive power.

As reported in the table, random forest clearly outperformed logistic regression in all tested combinations (see column AUC in Table 4). We further observed that in comparison to the logistic regression model, increasing the class weights for the HAD class reduced the AUC of the random forest. This might be because of the architecture of the random forest model that automatically reduces overfitting [30]. Furthermore, more involved sampling procedures performed better than the random ones (SMOTE outperformed random oversampling, EasyEnsemble outperformed random undersampling) and oversampling of the training data led to better results on the test data than undersampling. Altogether, random forest with SMOTE oversampling gave the best results with an AUC of 0.78.

The last set of experiments we performed concerned the feature ranking of the models. To find the features that played the most important role in the prediction, we ranked the feature importances of the best random forest model. This feature ranking is shown in Figure 2. The figure illustrates once again the importances of all features that measure the amount of help needed by the patients. The need for help how often is selected as the most important feature for the prediction task, followed by the general need for help and the need for supervision. In summary, also our final technique nominated the importance of the need for help and supervision features.

5. Conclusions

Prediction of the HAD syndrome has a great clinical value and understanding the underlying factors associated with HAD is a first step to better plan hospitalization of elderly patients. In particular, prevention and treatment of HAD consists of personally applied enhanced rehabilitation, which should be started early and be effective enough to prevent or treat the HAD. An enhanced rehabilitation is physically strenuous for the patient, and requires expertise and costs. Therefore, it is important to identify patients in risk of HAD soon after admitting into an hospital to start their rehabilitation as early as possible.

Although previous studies have tried to find factors associated with HAD (see Section 1), none of these have leveraged machine learning techniques. This paper presented a first attempt in this direction to automatically identify the relations of many health and social variables to the outcome of hospitalization of elderly patients. More precisely, using triangulation of different analysis techniques, it was shown that the need for help and supervision are the most important features predicting whether an elderly patient will return home after hospitalization due to an acute illness.

Our findings support earlier medical studies (see Section 1, especially Table 1). Indeed, the age of the patients itself was not found to be an as strong feature as pointed out in previous works (see, for example, [17, 2] described in Table 1 in comparison to Figure 2). However, our results confirm earlier studies that identified physical frailty and the ADL score at baseline as significant predictors for HAD (see, for example, [1, 19] described in Table 1). The worse the physical frailty and the ADL score at baseline, the more help and supervision the elderly patients needed already before admission to the hospital. Thus, our recommendation is to pay special attention to physical frail patients who already needed some help before their hospital admission and to provide them with special care and rehabilitation practice.

Future work will repeat the presented analysis
scheme for larger data from more hospitals. Although the random forest model outperformed the logistic regression model and SMOTE oversampling further improved the performance, an AUC of 0.78 is still quite distant from perfect classification. Thus, future work could benefit from a larger sample size of HAD patients for building models with refined AUC and sensitivity.

Finally, it would be interesting to exploit the potential of adding various secondary health care data to the model [35]. For example, current work [36, 37, 38] focuses on automatic summarizing methods for clinical free text notes. We expect that our model could be further enhanced if our raw data could be enriched by such high-level interpretations summarizing clinical free text notes describing the elderly patients. Another example of interesting secondary data is the increasing prevalence of self-taken or automated measurements at home that are accumulating personal health records (PHRs) in commercial and public data management systems. These PHRs may also provide relevant predictors to expedite the clinical decision making with HAD patients in the future. In Finland, for instance, all the citizens will be allowed to store a validated set of self-measurements into the national Patient Data Repository as of 2018.

Acknowledgments

This work has been carried out in two projects “Value from Public Health Data with Cognitive Computing” and “Watson Health Cloud”, funded by Business Finland.

References

Figure 2: Feature importances of the random forest model. The *need for help, how often* is the most important variable, followed by the general *need for help* and *need for supervision*. 