Decision Support Systems 147 (2021) 113539

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

Predicting employee absenteeism for cost effective interventions

Natalie Lawrance [a,*], George Petrides [a,b], Marie-Anne Guerry [a]

[a] Department of Business Technology and Operations, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium
[b] Department of Informatics, University of Bergen, Norway

ARTICLE INFO

Keywords:
Cost-sensitive learning
Classification
HR analytics
Absenteeism prediction

ABSTRACT

This paper describes a decision support system designed for a Belgian Human Resource (HR) and Well-Being Service Provider. Their goal is to improve health and well-being in the workplace, and to this end, the task is to identify groups of employees at risk of sickness absence who can then be targeted with interventions aiming to reduce or prevent absences. To facilitate deployment, we apply a range of existing machine-learning methods to obtain predictions at monthly intervals using real HR and payroll data that contains no health-related predictors. We model employee absence as a binary classification problem with loss asymmetry and conceptualise a misclassification cost matrix of employee sickness absence. Model performance is evaluated using cost-based metrics, which have intuitive interpretation. We also demonstrate how this problem can be approached when costs are unknown. The proposed flexible evaluation procedure is not restricted to a specific model or domain and can be applied to address other HR analytics questions when deployed. Our approach of considering a wider range of methods and cost-based performance evaluation is novel in the domain of absenteeism prediction.

1. Introduction

Employee *sickness absence* or *absenteeism*, broadly defined as failure to attend scheduled work as a result of ill health, is a pervasive problem disruptive to operations and costly to the economy. The annual cost of worker absenteeism in the countries of the Organisation for Economic Co-operation and Development (OECD) has been estimated to be between 1.2 and 2% of their total GDP [1], which in current terms translates to between 0.6 and 1 trillion US dollars [2]. It is therefore only natural for employers to seek solutions to this problem.

The motivation for this work was the request of a company in Belgium specialising in *Human Resource* (HR) Management and Well-Being for the development of a solution to address employee absenteeism using data science. A direction for finding a possible solution stemmed from the fact that as much as 20% of the working-age population in the OECD countries suffers from common mental illnesses such as anxiety and depression disorders [3], and timely application of preventive measures is crucial in avoiding transition to long-term illness and disability [1]. Interventions in the form of health management or wellness programs have a long history and several meta-analytic studies have reported strong evidence of their effectiveness at reducing employee absenteeism [4–6]. Examples of such interventions can include individual fitness program, stress-management seminars, private or group therapy sessions, and work flexibility arrangements, to name a few.

Naturally, such wellness programs are costly, both in terms of monetary costs for their implementation (which might not always be available due to budget limitation), and hours spent for participation. Therefore, the simplest solution of applying them to all employees at a workplace might not be the most cost-effective. Instead a design a decision support system is needed that is able to identify the employees at risk of sickness absence, typically only a small fraction of the workforce, who should then be targeted with a preventive action.

1.1. Related work

HR analytics is most commonly applied to talent acquisition [7] or retention [8], rather than maintaining and improving the well-being of employees. As observed in [9], the absenteeism prediction literature to date has mostly been concerned with the explanation and association of various risk factors with employee absenteeism, rather than accuracy of predictions. The majority of contributions to the domain of absenteeism prediction come from the field of occupational health and medicine [10–17]. In this field, statistical techniques such as logistic regression and Cox proportional-hazards models are preferred and training data is collected either in the clinical setting or as part of a nation-wide

[*] Corresponding author.
E-mail address: natalie.lawrance@vub.be (N. Lawrance).

https://doi.org/10.1016/j.dss.2021.113539
Received 8 August 2020; Received in revised form 22 February 2021; Accepted 24 February 2021
Available online 27 February 2021

household survey.

Another strand of literature focusses solely on predicting sickness absence using algorithmic models. Among those, several studies [18–20] approach employee sickness absence as a regression problem using neural networks that predict hours of absence on the data collected in [20]. [21] uses the same dataset in a multi-class classification setting using decision tree ensembles to predict absences of specific duration.

The main drawback in all of those studies is lack of transparency with regards to their experimental set-up and model selection procedure. Most contributions apply a single algorithm to a single dataset, the rationale behind the model choice and evaluation being rarely discussed. In contrast, in this paper we consider a wide range of state-of-the-art algorithms and relevant evaluation measures.

Previous studies involved data collection in several waves, usually annually. However, predictions with a horizon of one year are of little relevance to businesses that are trying to reduce direct and indirect costs incurred through loss of productivity and disruption to operations. What is more desirable is the ability of the management to obtain reliable predictions at operationally practical intervals (such as a month or a quarter) preferably with data that is readily available from HR and payroll records. An attempt has been made in [22] to predict sickness absence using only such data from one industry sector, albeit at the impractical one-year prediction horizon. The data we have been provided with for this work does not contain health-related or attitudinal information either, but includes several sectors and allows us to consider a more practical prediction horizon of one month.

Finally, as already mentioned, absenteeism is an event of relative rarity with only a small fraction of the population falling out of the workforce in any given period. Most absenteeism prediction papers do demonstrate an imbalanced class distribution of the outcome variable. Not accounting for this class imbalance in predictive models has immediate negative implications for model performance [23,24]. To the best of our knowledge, [25] is the only paper that recognises the importance of class imbalance correction and models sickness absence as a problem with loss asymmetry where cost of misclassifying an absentee is set to the number of non-absentees in the dataset, and vice-versa. This heuristic is commonly applied to treat class imbalance alone, but could be suboptimal when real costs can actually be specified. Specifying them is what we attempt to do to simultaneously address class imbalance and cost asymmetry using cost-sensitive learning.

1.2. Our contribution

In this work we investigate the use of predictive analytics as a decision support system for increasing employee well-being in the workplace by identifying groups of employees at risk of sickness absence that should be targeted with a wellness intervention.

Firstly, we employ cost-sensitive learning to treat the unequal misclassification costs pertaining absenteeism, something we could not find in the existing literature. Our main contribution is the conceptualisation of a relevant misclassification cost matrix, which can be generalised to a variety of institutional and legislative contexts, and consequently to other datasets as well. A core element of our cost matrix is information on the effectiveness an intervention has on individuals, which is currently lacking and therefore identified as an important direction for future research and requires collaboration between academia and industry. We also develop business-friendly cost-based evaluation metrics that have an intuitive interpretation.

Secondly, since our data did not contain information on one of the parameters of our cost matrix, we also consider a cost-insensitive approach. We evaluate performance using balanced accuracy, which assumes that misclassification errors have equal severity, despite the class imbalance.

Finally, we try to illustrate best practices in the domain of absenteeism prediction for the practitioner. For this, we use an anonymised dataset that a Belgian HR and Well-Being company provided us, which contains employee payroll information only, without any health-related data. We specify different cost-matrices by considering different realistic scenarios for interventions and their (fictional) effectiveness, and consider a practical prediction horizon of one month. We follow a rigorous experimental design to prevent over-fitting and develop a flexible algorithm-agnostic evaluation framework. In the end, a selection of tree-based classifiers, both cost-sensitive and cost-insensitive, is evaluated on the same cost-matrices, and the advantages and disadvantages of both approaches are discussed.

2. Preliminaries

In what follows we will briefly outline the challenges of classification on imbalanced data. Readers familiar with the material are invited to continue to Section 3.

A classifier is a function $f : x \to \hat{y}$ that maps a vector of real-valued predictors $x \in \mathbb{R}^n$ to a predefined target class $\hat{y} \in \mathbb{R}$ based on a training set of data with known true class labels $y \in \mathbb{R}$. In a binary classification setting, it is common to consider class labels $y, \hat{y} \in \{0, 1\}$, with the rare class referred to as positive and labelled as 1.

Most binary classifiers produce predictions in two stages: First a confidence $score \in [0, 1]$ is produced for each observation. Then, an instance i is classified as positive ($\hat{y}_i = 1$) if its score is greater than a threshold T, and as negative otherwise. Most classifiers implicitly assume $T = 0.5$, which often results in poor classifier performance under class imbalance [23,26].

The typical loss function adopted to measure a classifier's prediction accuracy is the 0–1 loss function, which counts the instances of incorrect classification. More detailed error analysis can be conducted using a confusion matrix, an example of which is shown in Table 1. Each entry in the confusion matrix represents the number of observations in the test set that were classified either correctly or incorrectly. The error count is thus split between two error types: false positives and false negatives.

2.1. Cost-sensitive decision-making

In most domains different misclassification errors entail different costs, known as cost asymmetry or sometimes cost skew or imbalance, and therefore the question of which of the error types is more costly is determined by the area of application. When the problem under consideration suggests cost asymmetry a cost-sensitive classification approach becomes appropriate [23]. Cost-sensitive learning translates the error-minimisation problem to cost-minimisation, where each prediction type (as defined in the confusion matrix) is assigned a cost by means of a misclassification cost matrix $\begin{pmatrix} C_{TP} & C_{FN} \\ C_{FP} & C_{TN} \end{pmatrix}$. This matrix can be either class-dependent, where all observations of a class entail identical costs, or they can be record-dependent in which case every observation i has its own cost matrix $\begin{pmatrix} C^i_{TP} & C^i_{FN} \\ C^i_{FP} & C^i_{TN} \end{pmatrix}$, derived from a given dataset.

An important result derived by Elkan [27] is that in the case where misclassification costs are known, any classifier can be made cost-sensitive by adopting a decision threshold that incorporates these costs, a method referred to as *Direct Minimum Expected Cost Classification (DMECC)* [28,29]. Here, the decision threshold is defined as follows:

Table 1
Confusion matrix. Each entry represents the number of observations in the respective category on a given test set.

	Predicted 1	Predicted 0
True 1	True Positive (TP)	False Negative (FN)
True 0	False Positive (FP)	True Negative (TN)

$T_{cs} = \frac{C_{FP}-C_{TN}}{C_{FP}-C_{TN}+C_{FN}-C_{TP}}$, where *cs* stands for cost-sensitive and the threshold can be either class- or record-dependent. This expression can be simplified to $T_{cs} = \frac{C'_{FP}}{C'_{FP}+C'_{FN}}$ if we transform the cost matrix into $\begin{pmatrix} C'_{TP} = 0 & C'_{FN} = C_{FN} - C_{TP} \\ C'_{FP} = C_{FP} - C_{TN} & C'_{TN} = 0 \end{pmatrix}$.

Another way to make any classifier cost-sensitive is to use an approach called *thresholding* [30], which performs a search across all scores produced by a given classifier. The score giving the lowest cost-loss is chosen as the decision threshold. In a situation where costs are unknown, *thresholding* can be optimised using a suitable loss-metric instead to improve classification performance on imbalanced datasets [31].

2.2. Classifier performance evaluation under cost asymmetry

The goal of cost-sensitive learning is to construct a classifier that is aware of the differences in importance between the classes. The advantage of using a cost-matrix is the exact specification of the loss function for any given data input. This, of course, calls for a suitable performance metric. In the cost-sensitive literature classifier performance is typically measured in terms of the total expected misclassification cost [27,32], which is simply the total cost-weighted classification error. Let S_{FP} be the set of false positives produced by a given classifier on a given test set, and let S_{FN} be the set of false negatives. The total misclassification cost of a given classifier is $TC = \sum_{i\in S_{FP}} C^{i}_{FP} + \sum_{i\in S_{FN}} C^{i}_{FN}$. However, how must one evaluate a classifier when the costs are unknown at the time of estimation? Here we again refer to the importance of the knowledge of the application domain, which can determine whether or not performance on one of the two classes should be favoured. In the domain where both types of misclassification costs are non-negligible, it is preferable to use a metric that incorporates performance with regard to both classes, rather than one that only favours the positive class. Many metrics exist that assess classifier performance [33,34], with most derived from the confusion matrix (see Table 1).

An empirical study of the stability of several such performance metrics under various degrees of class skew concludes that the two metrics that remain unbiased in the presence of class skew are the true positive rate $TPR = \frac{TP}{P}$ and the true negative rate $TNR = \frac{TN}{N}$. Their arithmetic average $\frac{TPR+TNR}{2}$ shares these desirable properties [35]. This metric places equal emphasis on each misclassification error type, which, in the absence of information regarding the importance of each of the two classes, is a reasonable choice. It is known under several names in the literature, such as balanced accuracy (*BACC*) [36,37], bookmaker informedness [35] or weighted accuracy [33] and happens to correspond to a point on the receiver operating characteristic curve (ROC) at a given decision threshold [33,35,37]. In this paper, we also consider the cost-effectiveness of the best models selected using *BACC* in case the costs were known.

3. A cost matrix for employee absenteeism and well-being interventions

In this section we present our conceptualisation of a cost matrix of the direct costs of employee sickness absence in relation to a well-being intervention. We consider this as one of the main contributions of this paper.

In any period M, an employee is contractually obligated to supply t_M hours of work in return for remuneration W, yielding the *base hourly rate* $\frac{W}{t_M}$ of the employee for this period. If in this period the employee is absent due to sickness for a total duration of $t_s \in [0, t_M]$ hours, the number of worked hours is reduced to $t_M - t_s$, which are remunerated as usual according to the base rate. However, depending on the legislation of the country of employment, the employer may also be required to remunerate the t_s hours of sickness according to a proportion $r \in [0,1]$ of the employee's base rate[1]. The hourly rate is in this case equal to $\frac{(t_M-t_s)\frac{W}{t_M}+t_s r\frac{W}{t_M}}{t_M-t_s} = \left(1+\frac{t_s r}{t_M-t_s}\right)\frac{W}{t_M}$, which is higher than the base rate, reflecting the loss of productivity associated with the employee's absence.

Suppose now that the employer decides to put the employee through a well-being intervention in an attempt to prevent potential sickness absence. The price of such an intervention per participant is C, which burdens the employer. In addition, if the intervention requires attendance (such as a coaching seminar) of duration t_i, the number of hours worked by the employee is reduced by as much. Therefore, the hourly rate of an employee who was not going to be absent but is put through an intervention is $\frac{W+C}{t_M-t_i}$, which is also higher than the base rate. If, however, the employee was going to be absent, in addition we expect that the intervention would have a positive effect and result in the reduction of the absence period by $\widetilde{t}_s \in [0, t_s]$ hours[2]. The resulting hourly rate is then equal to

$$\frac{(t_M - t_s + \widetilde{t}_s)\frac{W}{t_M} + (t_s - \widetilde{t}_s)r\frac{W}{t_M} + C}{t_M - t_s - t_i + \widetilde{t}_s} = \frac{W+C}{t_M - t_s - t_i + \widetilde{t}_s} - \frac{\left(t_s - \widetilde{t}_s\right)(1-r)}{t_M - t_s - t_i + \widetilde{t}_s}\frac{W}{t_M}$$

. Table 2 summarises these hourly rates. Here, just like in the confusion matrix in Table 1, the rows correspond to the true outcomes, and columns correspond to predicted outcomes. Thus, true positives are absentees targeted with an intervention, false positives are non-absentees targeted with an intervention, true negatives are non-absentees not targeted, and finally, absentees not targeted are false negatives.

Remark 1. Clearly, a necessary condition for the intervention to be cost-effective for the employer is that $t_i < \widetilde{t}_s$.

Remark 2. A limiting factor in specifying a concrete cost matrix is the parameter $\widetilde{t}_s$, which is a priori unknown and no indication of example values can be found in the literature.

3.1. The case of Belgium

As we mentioned in the introduction, this work initiated at the request of a Belgian HR and Well-Being Specialist. We therefore adapt Table 2 to the specifics of Belgian legislation.

In Belgium, throughout all sickness absences lasting up to 30 calendar days, white-collar workers receive full wage equivalent sickness benefits from the employer. As soon as the duration of absence is longer than one calendar month, the benefits are paid by the social security instead. Blue-collar workers receive reduced compensation starting from week two of absence: the employer continues to cover some fraction r of the full wage W, while the remainder is covered by social security [38].

Our sample contains only white-collar employees. By consequence, the parameter $r = 1$ and t_s is defined as the total hours of sickness absences covered by the employer in any given month M, with the necessary condition that $t_s < t_M$.

Using the hourly rates from Table 2, and after applying the transformation as mentioned in Section 2.1, such that the cost of correctly classifying observations is zero, we obtain the following costs:

[1] In some countries, r may actually be non-constant but instead a function of t_s. Our arguments can be extended to such case as well, but for the sake of clarity here we focus on constant r.

[2] For simplicity we assume that $\widetilde{t}_s$ cannot be observed beyond the prediction horizon of one period ahead, and we leave the dynamic case to future work.

Table 2
The cost matrix of employee sickness absence in terms of hourly rates of employee remuneration, when considering well-being intervention. W represents the employee's salary, t_M the expected work hours, t_s is the absence duration in hours, r is the fraction of the base rate $\frac{W}{t_M}$ to which hours of absence are remunerated as guaranteed by law in the form of statutory sick pay, $\widetilde{t}_s$ is the reduction in hours of sickness absence because of the intervention, C is the cost of the intervention, and t_i is the duration of attendance in hours associated with the intervention.

	Intervention	No Intervention
Absentee	$C_{TP} = \frac{W+C}{t_M - t_s - t_i + \widetilde{t}_s} - \frac{(t_s - \widetilde{t}_s)(1-r)}{t_M - t_s - t_i + \widetilde{t}_s} \frac{W}{t_M}$	$C_{FN} = \left(1 + \frac{t_s r}{t_M - t_s}\right) \frac{W}{t_M}$
Non-Absentee	$C_{FP} = \frac{W+C}{t_M - t_i}$	$C_{TN} = \frac{W}{t_M}$

$$\begin{pmatrix} C'_{TP} = 0 & C'_{FN} = \frac{W}{t_M - t_s} - \frac{W+C}{t_M - t_s + \widetilde{t}_s - t_i} \\ C'_{FP} = \frac{W+C}{t_M - t_i} - \frac{W}{t_M} & C'_{TN} = 0 \end{pmatrix} \tag{1}$$

To simplify notation we omit the superscripts i that indicate that costs are record-dependent.

Remark 3. Note that while C'_{FP} is always positive, C'_{FN} may become negative when $t_i > t_s \geq \widetilde{t}_s$. This violates the so called reasonableness condition defined by Elkan [27], which states that the cost of correct predictions should always be less than the cost of misclassifying, otherwise it is more profitable to misclassify than to classify correctly. A negative misclassification cost is a benefit for the employer, meaning that it is more cost-effective not to apply an intervention to that individual (e.g. due to low expected sickness absence hours).

4. Experimental framework

In this section we present the experimental procedure that was used to conduct our analysis. Our experiments consist of two parts: in the first instance we predict employee sickness absence using cost-sensitive learning and the cost matrix defined in Eq. (1), we evaluate model performance using custom cost-based metrics. Then, in view of lack of data regarding the $\widetilde{t}_s$ parameter, we additionally predict using cost-insensitive models, which we evaluate using the balanced accuracy score. In both cases, we also report standard metrics such as the *AUC, FPR, FNR*.

4.1. Data

Our data contains HR and payroll records from roughly 280 small, medium and large Belgian firms from a variety of industry sectors. The data spans the period between January 2018 and March 2019. We adopt the prospective study design so commonly found in absenteeism prediction literature, where attributes from period M_t are used to predict the outcome in period M_{t+1}.

4.1.1. Target variable

In any given month, each employee's hours of certified sickness absences[3] are summed and converted to a binary target variable according to a threshold T_{hrs} : observations having a total number of hours recorded below T_{hrs} are coded as 0, and the rest as 1. The choice of T_{hrs} should of course depend on the task at hand. In our case, after consulting the data provider, we decided to set $T_{hrs} = 0$. The resulting distribution of our target variable is highly imbalanced, ranging between 7.5% and 16.5% of positives in any given period. Table 3 shows the class imbalance in our data per prediction period.

4.1.2. Predictors

One of the novelties of our work is to exploit the rich absence pattern data at our disposal. The main difficulty we are faced with is the absence of any health-related predictors in our dataset or the reason for absence. Our dataset consists of 66 features.

Demographic features: Employee's demographic features, such as age, gender, marital status, education etc.

Work environment features: Features that describe various aspects of the work circumstances of employees (wage, contract type, etc.). We also include fatigue inducing factors such as work shift irregularities (e.g. weekend work, overtime, night shifts) as well as patterns of holiday applications (e.g. holiday frequency and duration, number of rejected holiday applications, time since last holiday of a certain duration).

Historic absence patterns: measures of recency (time since last absence) and frequency of illnesses, average hours of sickness absences in the 12 months prior to the prediction period and since the start of employment contract.

4.1.3. Data preparation

In continuous numeric variables all values exceeding plausible minima and maxima are removed (e.g. age values below 18 and above 100) and missing values are imputed with sample median. All levels in categorical predictors are transformed to binary variables, including the missing values. In recency variables missing values are replaced with 366 indicating that the last incident was registered more than one calendar year ago.

4.2. Methods

4.2.1. Classification algorithms

We adopt a wide range of decision tree ensembles in our framework and combine them with state-of-the-art solutions to the problem of class imbalance. Since a requirement for our decision support system was that it could be readily implemented in the cloud, we apply methods with existing open-source implementations.

The base algorithm adopted in our experiments is a decision tree classifier (specifically CART [39]). Decision trees have been shown to have a number of highly desirable properties: they can handle mixed data types and missing values; they are insensitive to monotone transformations of the feature space and do not require normalisation of predictors; they can handle irrelevant predictors and are robust to

Table 3
The class imbalance in our data per prediction period (year/month).

	Period Attributes	Period Target	# employee's (from # firms)	# positives (%)
1	2018/01	2018/02	50,729 (284)	8193 (16.15)
2	2018/02	2018/03	49,459 (281)	8161 (16.50)
3	2018/03	2018/04	48,202 (280)	4268 (8.85)
4	2018/04	2018/05	51,998 (280)	4682 (9.00)
5	2018/05	2018/06	51,483 (280)	5123 (9.95)
6	2018/06	2018/07	50,843 (281)	3845 (7.56)
7	2018/07	2018/08	51,372 (282)	3863 (7.52)
8	2018/08	2018/09	51,738 (282)	5201 (10.05)
9	2018/09	2018/10	51,130 (282)	6340 (12.40)
10	2018/10	2018/11	50,744 (281)	5905 (11.64)
11	2018/11	2018/12	49,659 (268)	4986 (10.04)
12	2018/12	2019/01	47,128 (268)	6256 (13.27)
13	2019/01	2019/02	52,637 (268)	8432 (16.02)
14	2019/02	2019/03	51,751 (268)	6493 (12.55)

[3] The medical reason for absence is unknown and therefore all types of absences are treated equally. This limitation is addressed in Section 6.1

outliers [40]. These properties combined with model interpretability make decision trees highly suitable for business applications. While a single decision tree may not show the highest performance because of high variance, combining a collection of trees into an ensemble decreases variance and improves performance, though at the expense of reduced interpretability [41].

All of the algorithms applied in our experiments are cost-insensitive by default, but can be made cost-sensitive in the presence of explicitly defined costs. In this work we limit ourselves to pre-training and post-training cost-sensitive learning methods as categorised in [42]. We describe these methods below and Table 4 provides a summary.

4.2.1.1. Class imbalance correction. Each of the cost-insensitive decision tree ensembles can be combined with a sampling method in order to compensate for unequal class distribution of the training data. This can be achieved by modifying the training set using cost-sensitive sampling and passing it to a classifier of choice (also referred to as *CS-preSampleEnsemble*). Alternatively resampling can be performed at the level of the ensemble (also referred to as *CS-SampleEnsemble*), where both classes are sampled in appropriate proportions prior to training each of the base classifiers in the ensemble.

We performed cost-sensitive data sampling using average training misclassification costs, defined as follows: $\overline{C}_{FP} = \frac{1}{N}\sum_{i \in S_N} C^i_{FP}$, $\overline{C}_{FN} = \frac{1}{P}\sum_{i \in S_P} C^i_{FN}$. Here S_N is the set of negatives (of size N) and S_P is the set of positives (of size P). The cost ratio $\frac{\overline{C}_{FN}}{\overline{C}_{FP}}$ is used to scale the number of positives for over-sampling and $\frac{\overline{C}_{FP}}{\overline{C}_{FN}}$ is used to scale the number of negatives for under-sampling [42].

When costs are unknown, it is reasonable to assume that the optimal class distribution is uniform and the classes are sampled in equal proportions [27].

Table 4
Overview of the methods used in our analysis.

Our notation	Method	Parameter settings
	Algorithms	
dt	Decision Tree (CART [39])	
	Ensembles	
bag	Bagging [43]	#trees = 100
rdf	Random Decision Forest [44]	#trees = 100
rf	Random Forest [45]	#trees = 100
xrf	Extremely Randomized Trees [46]	#trees = 100
adab1	AdaBoost [47]	#nodes = 1, #trees = 50
adab2		#nodes = 2, #trees = 50
CS-WeightedEnsemble		
wrf	Weighted Random Forest [48]	#trees = 100
CS-SampleEnsemble		
bal_rf	Balanced Random Forest [48]	#trees = 100
bal_rdf	Balanced Random Decision Forest	#trees = 100
easy_ensmb	Easy Ensemble [49]	#estimators = 50
CS-SampleBoost Methods		
rusboost	RUSBoost [50]	#trees = 50
	Sampling Methods	
ros	Random Oversampling	
rus	Random Undersampling	
smote	SMOTE [51]	$k = 5$
	Calibration	
isotonic	Isotonic Regression [52]	
sigmoid	Platt [53]	
	Post-training methods	
$T = T_{best}$	Thresholding [30]	
$T = T^i_{cs}$	DMECC [29]	

4.2.1.2. Calibration. As was observed in [29,54], decision tree methods do not produce reliable posterior class membership probability estimates. This has negative implications for application of post-training methods such as described above, where class membership probabilities are used for decision-making [52]. In this work, we optionally apply two probability calibration methods to each of the classifiers in our framework: isotonic regression [52] and Platt scaling [53].

4.2.1.3. Post-training methods. When costs are record-dependent and are known at the time of estimation *DMECC* can be applied to derive the total misclassification cost on a given test set. If costs are unknown, *thresholding* can be applied to optimise over an error-based loss, which assumes that costs are class-dependent.

All of our experiments consider the default threshold $T = 0.5$ along with the post-training methods that are simple to implement in practice.

Cost-sensitive models are evaluated at the decision threshold obtained using the *DMECC* approach (explained in Section 2.1). For the application of *DMECC*, both types of misclassification costs from Eq. (1) need to be specified to calculate record-dependent decision thresholds T^i_{cs} . Since t_s is unknown at the time of prediction, we estimate this in two ways. First, we set the t_s equal to each record's individual average sickness duration in 12 months prior to prediction (denoting the resulting model $T^i_{cs} - mean$). Second, we predict t_s using a Random Forest regressor trained on positive records from the same training set that the cost-sensitive classifier is trained on (denoted $T^i_{cs} - postreg$). Cost-insensitive models classify at the optimal decision threshold $T = T_{best}$ obtained using *thresholding* (see Section 2).

All of the above experiments were run on the same 13 pairs of training and test data periods. In our experiments we use the combination of twelve algorithms, four pre-training options (including no pre-training), three calibration options (including no calibration), which gives us 120 base models to which post-training was optionally applied. Cost-insensitive models were evaluated at two decision thresholds, resulting in 240 models per period. From 120 base models 111 could be made cost-sensitive, each of which were evaluated at three decision thresholds (explained in 4.2.1), resulting in 333 models per period and per combination of C and t_i, and $\widetilde{t_s}$ (described in 4.4.2). CS-SampleEnsemble and CS-SampleBoost methods, naturally, are only combined with calibration. We note that using the name Balanced Random Forest in both cost-sensitive and cost-insensitive models we abuse the terminology slightly, as in fact cost-sensitive models perform sampling according to the cost matrix.

4.3. Software used

All experiments were conducted using Python (version 3.6.6), the majority of implementations come from *scikit-learn* (version 0.20.2) library [55], implementations of all pre-training methods come from *imbalanced-learn* library [56], post-training methods are our own implementations; the results were processed using Python and *csvkit* [57].

4.4. Experimental set up

4.4.1. Data partition

Three datasets, that contain the same set of employees, served as input to the experimental framework. We randomly split the set of employees into three disjoint sets. The first contains 60% of all observations and was used for training. The second contains 20% of the total, used for optimal threshold search respectively, where features were from period M_t and labels were from M_{t+1}. The remaining 20% of employees were used for testing with features from M_{t+1} and labels from M_{t+2}. This process was repeated 50 times using different random seeds to generate data splits and the results were averaged across iterations.

4.4.2. *Training costs*

The definition of the cost parameters are as introduced in Section 3.

The values for W, t_M and t_s are readily available from our data. In the absence of information regarding the effectiveness of chosen interventions in reduction of sickness absence, we assume some arbitrary values for the parameter $\widetilde{t_s}$. We consider three alternatives for this parameter: a) $\widetilde{t_s} = t_s$, b) $\widetilde{t_s} = 0.5 \cdot t_s$, and c) $\widetilde{t}_s^i = p^i \cdot t_s^i$ where $p^i \in [0, 1]$ is an individual percentage of reduction in sickness hours, randomly drawn from a uniform distribution. This parameter was drawn once and was reused across all intervention scenarios. We expect that alternative c) reflects the reality best, but further experiments are needed (see Section 6.1).

We consider three examples of interventions with corresponding price C and duration t_i :

Case 1. "Fit Check-Up". Aimed at increasing physical activity levels of employees. A professional coach examines one's fitness level using specialised equipment and designs a personalised 12-week fitness programme. The parameter settings are as follows: $C = 100$(EUR), $t_i = 4$ (hours).

Case 2. A sleep and fitness tracking device. Aimed at promoting healthy lifestyle choices. The built-in software informs on sleep and activity patterns and encourages participants towards positive change in behaviour. The parameter settings are as follows: $C = 30$(EUR), $t_i = 0$ (hours).

Case 3. Psychotherapy. The employer offers financial support for individual psychotherapy sessions to help reduce stress and prevent burnout. The parameter settings are as follows: $C = 60$(EUR), $t_i = 1$(hour).

4.4.3. *Evaluation*

Cost-sensitive model performance is typically assessed using the model's total misclassification cost (*TC*), which sums record-dependent costs on a given test set. In the absence of a decision support system, the employer has two naive solutions, namely not applying any intervention, and applying the intervention indiscriminately to all employees, with respective costs TC_{none} and TC_{all}. Any intervention target group proposed by a predictive model (henceforth referred to as the campaign) is worthwhile only if it improves on either of these naive costs, something we are trying to capture via the *Cost Improvement Score*: $CIS = 1 - \frac{TC}{min(TC_{none}, TC_{all})}$, with the following interpretation. $CIS < 0$ indicates that the model's intervention campaign is more costly than either of the naive approaches. A perfect model that correctly classifies everyone achieves $CIS = 1$. To evaluate whether or not the size of the model's intervention campaign prescribed by the model is cost-effective, as well as to compare models with similar cost performance, we use *Return On Investment* (*ROI*). In the context of absence prediction, *ROI* is defined as the intervention profit (benefits minus costs) over intervention costs:

$$ROI = \sum_{j \in S_{PP}} \frac{\widetilde{t}_s^j \frac{W^j}{t_M^j}}{C + t_i \frac{W^j}{t_M^j}} - 1$$

, where S_{PP} is the set of positives predicted by the model and the index j refers to the fact that both the costs and the profits are record-dependent. $ROI < 0$ indicates that the model's intervention campaign prescribed by the model is not cost-effective. Both *CIS* and *ROI* should be considered in the final model selection. For example, when one model has $CIS = 0.7$ and *ROI* is some positive value A, and another model has $CIS = 0.68$ with $ROI\ B > A$, then the second model might be preferred as it requires less budget to be available for similar cost performance.

5. Results and discussion

The primary interest of our work lies in applying cost-sensitive learning to the problem of absenteeism prediction on real data, using our newly designed cost matrix. We evaluate model performance per period, based on the two cost metrics: *CIS* and *ROI* (defined in Section 4.4.3). Related to this are two further questions: first, whether the usage of record-dependent costs offers any advantage over class-dependent costs, and second, whether predicting the value of t_s using a regressor offers any advantage over using mean historic sickness per individual. We also discuss which of the interventions considered offers the lowest misclassification cost. With the given current difficulty in obtaining all the necessary parameters of the cost matrix, we also try to see whether selecting models using a cost-insensitive metric would be a good alternative to cost-sensitive learning.

Analysing the performance of specific algorithms lies outside the scope of this paper.

5.1. *Cost-sensitive absenteeism prediction*

The first part of our experiments provides an illustration of how our proposed misclassification cost matrix of employee absenteeism (defined in Eq. (1)) performs on real world data.

5.1.1. *Model performance and cost-effectiveness*

Our results show that in every prediction period and intervention combination, we find cost-sensitive models with $CIS > 0$, and with $ROI > 0$.

Fig. 1 demonstrates cost performance of the top-ranking models (when ranked by *CIS*). We note that the largest cost improvement (the difference between the benchmark and the total cost) of the model was achieved under Case 1 and Case 3, while Case 2 - the least expensive intervention - shows only marginal improvement. We also found that *DMECC* models that use regression predictions to classify, achieved higher cost improvement much more frequently than models that use average hours.

5.1.1.1. Record-versus class-based costs. As was mentioned in Remark 3, in our cost matrix, the reasonableness condition only holds for those individuals j, whose $\widetilde{t_s}^j > t_i$, i.e. when the expected reduction in hours of absence due to the intervention exceeds the duration of the intervention. The *DMECC* approach can directly account for this condition, which allow us to avoid targeting individuals for whom an intervention does not pay off. Models that use a constant threshold for classification decision, do not make such a distinction, and are less frequently ranked among the top models.

Another drawback of using class-based costs became apparent when we applied sampling methods in combination with averaged record-dependent costs (explained in Section 4.2.1). We observed that under certain operating conditions, for a number of employees in our sample holds that $C_{FP}^i > C_{FN}^i$, and when the proportion of such individuals in the training sample was large enough, this resulted in: $\overline{C}_{FP} > \overline{C}_{FN}$. Under such cost imbalance, an oversampler will oversample the more costly class i.e. the negatives and an undersampler will undersample the less costly class i.e. the positives, which would increase class imbalance, instead of correcting it. Such models failed in the pre-training phase and were not considered in the rankings.

5.1.2. *Cost-effectiveness of interventions*

Another question a practitioner might be interested in is: which intervention offers the highest savings in any given period? To determine this, we shortlist the intervention that yield the highest *CIS* on any given prediction period and $\widetilde{t_s}$ combination. These results are presented in Table 5. We note that the highest *CIS* is not necessarily associated with a positive *ROI*, and therefore we restrict ourselves to models that have

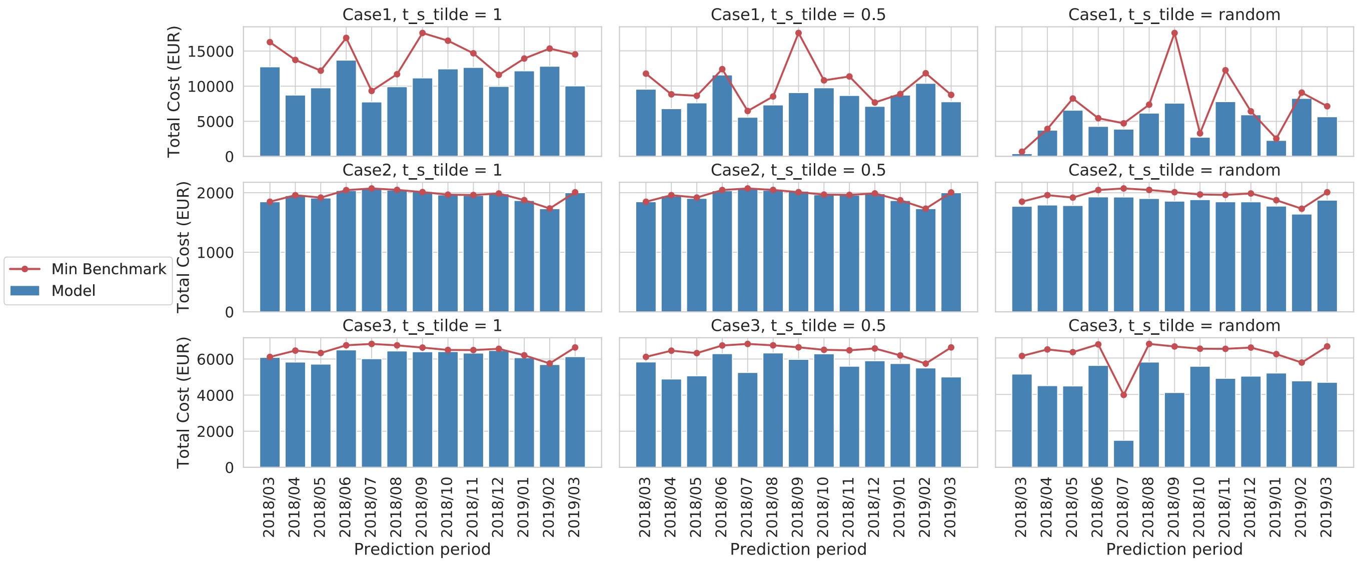

Fig. 1. Cost performance of the cost-sensitive models across all prediction periods and interventions. The bars represent the total misclassification cost of the top-ranking model (ranked by *CIS*). The solid line shows the cost of the benchmark under the same operating conditions. Here "t_s_tilde" is $\widetilde{t}_s/t_s$ and "random" refers to record-dependent $p_i \in [0, 1]$.

Table 5
Selecting the most cost-effective intervention. For the final decision of which of the interventions offers the highest cost improvement in any given prediction period, we have selected models with the highest *CIS* and positive *ROI*. The column titled "% pos with $C_{FN} < 0$" shows the percentage of individuals in the test set whose misclassification costs violate the reasonableness conditions (see Remark 3). Rand refers to record-dependent $p_i \in [0,1]$.

Target period	Intervention		Method	%pos	%pos with $C'_{FN} < 0$	Error-rates		Cost-based metrics			
	$\widetilde{t}_s/t_s$	Case	algo_sampling_calib-T			FNR	FPR	CIS	TC	Benchmark	ROI
2018/03	1	1	xrf_rus_sigmoid-T^i_{cs}-postreg	15.62	7.94	0.34	0.46	0.21	12,756.80	16,164.63	0.21
2018/03	0.5	1	easy_ensmb_none_none-$T = 0.5$	15.62	26.59	0.59	0.19	0.18	9545.19	11,695.74	0.41
2018/03	rand	1	dt_ros-isotonic-T^i_{cs}-postreg	15.62	40.61	0.81	0.13	0.98	384.20	16,164.63	1.01
2018/04	1	1	adab1_smote_sigmoid-T^i_{cs}-postreg	8.09	12.43	0.54	0.27	0.32	8733.03	12,789.16	0.07
2018/04	0.5	1	xrf_ros-isotonic-T^i_{cs}-postreg	8.09	33.34	0.84	0.09	0.26	6777.96	9218.97	0.48
2018/04	rand	1	wxrf_ros_sigmoid-T^i_{cs}-mean	8.09	43.55	0.98	0.01	0.55	3735.29	8297.38	13.81
2018/05	1	1	adab2_rus-isotonic-T^i_{cs}-postreg	8.01	12.51	0.82	0.07	0.26	9765.23	13,119.91	2.39
2018/05	0.5	1	rusboost_none_none-T^i_{cs}-mean	8.01	36.65	0.96	0.02	0.24	7596.42	9939.64	3.77
2018/05	rand	1	bal_rf_none_sigmoid-T^i_{cs}-postreg	8.01	48.01	0.91	0.02	0.30	6588.44	9386.79	3.49
2018/06	1	1	adab2_ros_sigmoid-T^i_{cs}-postreg	9.05	13.88	0.80	0.06	0.23	13,708.92	17,892.66	3.52
2018/06	0.5	1	adab2_none_sigmoid-T^i_{cs}-postreg	9.05	37.53	0.91	0.01	0.17	11,569.85	13,893.63	6.73
2018/06	rand	1	adab1_none-isotonic-T^i_{cs}-postreg	9.05	47.53	0.92	0.02	0.67	4281.80	13,173.46	6.24
2018/07	1	1	adab2_rus-isotonic-T^i_{cs}-mean	6.87	13.34	0.89	0.03	0.20	7750.51	9747.05	5.96
2018/07	0.5	1	rf_ros_none-T^i_{cs}-mean	6.87	35.85	0.95	0.01	0.18	5556.31	9747.05	13.39
2018/07	rand	3	xrf_smote-isotonic-T^i_{cs}-postreg	6.87	20.19	0.59	0.27	0.78	1481.94	6828.25	0.11
2018/08	1	1	xrf_none_none-T^i_{cs}-postreg	6.72	14.15	0.79	0.07	0.15	9918.14	11,708.51	1.90
2018/08	0.5	1	wrf_none_sigmoid-T^i_{cs}-postreg	6.72	37.27	0.94	0.01	0.16	7306.00	8700.07	11.39
2018/08	rand	1	bal_rf_none-isotonic-T^i_{cs}-postreg	6.72	47.90	0.87	0.02	0.23	6174.96	8053.30	3.75
2018/09	1	1	easy_ensmb_none_none-$T = 0.5$	9.28	11.76	0.50	0.24	0.36	11,172.95	17,555.27	0.45
2018/09	0.5	1	bal_rf_none-isotonic-$T = 0.5$	9.28	33.78	0.83	0.05	0.48	9062.25	17,555.27	2.36
2018/09	rand	1	bal_rf_none-isotonic-T^i_{cs}-postreg	9.28	45.52	0.86	0.04	0.57	7594.42	17,555.27	2.54
2018/10	1	1	adab1_rus_sigmoid-T^i_{cs}-postreg	11.57	11.27	0.45	0.33	0.26	12,463.05	16,772.53	0.21
2018/10	0.5	1	wrf_none_none-T^i_{cs}-postreg	11.57	32.20	0.92	0.04	0.15	9743.45	11,519.75	3.46
2018/10	rand	1	rdf_smote_sigmoid-T^i_{cs}-postreg	11.57	44.39	0.91	0.04	0.71	2729.36	9273.33	4.47
2018/11	1	1	xrf_smote_none-T^i_{cs}-postreg	10.86	10.97	0.65	0.20	0.16	12,671.80	15,114.80	0.78
2018/11	0.5	1	easy_ensmb_none_none-T^i_{cs}-mean	10.86	32.09	0.95	0.03	0.17	8649.31	10,381.59	2.52
2018/11	rand	1	adab2_none-isotonic-T^i_{cs}-postreg	10.86	43.67	0.88	0.05	0.07	8496.65	9133.13	2.35
2018/12	1	1	easy_ensmb_none_sigmoid-T^i_{cs}-postreg	8.95	11.37	0.72	0.12	0.17	9969.19	12,006.99	1.47
2018/12	0.5	1	wrf_none-isotonic-T^i_{cs}-postreg	8.95	31.98	0.90	0.02	0.11	7109.03	7964.33	5.29
2018/12	rand	1	rf_none_sigmoid-T^i_{cs}-postreg	8.95	45.67	0.92	0.00	0.15	5932.89	6994.47	6.19
2019/01	1	1	wrf_rus_none-T^i_{cs}-postreg	12.37	9.36	0.65	0.24	0.14	12,184.36	14,212.21	0.64
2019/01	0.5	1	rdf_none_sigmoid-T^i_{cs}-postreg	12.37	30.17	0.94	0.02	0.04	8707.68	9084.91	9.15
2019/01	rand	1	adab1_none-isotonic-T^i_{cs}-postreg	12.37	42.44	0.95	0.02	0.72	2282.78	8058.50	7.78
2019/02	1	1	adab1_rus_sigmoid-T^i_{cs}-postreg	14.69	10.41	0.52	0.26	0.16	12,839.96	15,316.98	0.83
2019/02	0.5	1	bal_rf_none_sigmoid-T^i_{cs}-postreg	14.69	31.72	0.79	0.07	0.08	10,378.46	11,346.17	2.26
2019/02	rand	1	adab2_none_sigmoid-T^i_{cs}-postreg	14.69	44.57	0.84	0.06	0.19	8279.77	10,169.78	2.45
2019/03	1	1	easy_ensmb_none-isotonic-T^i_{cs}-postreg	11.62	12.17	0.58	0.19	0.32	10,038.56	14,837.82	0.89
2019/03	0.5	3	adab1_ros_sigmoid-T^i_{cs}-postreg	11.62	6.69	0.36	0.39	0.25	5005.78	6637.93	0.23
2019/03	rand	1	adab2_rus-isotonic-T^i_{cs}-postreg	11.62	47.45	0.83	0.08	0.32	5653.64	8300.72	1.20

both $ROI > 0$ and $CIS > 0$.

5.2. Absenteeism prediction when costs are unknown

As mentioned in Remark 2, the value of the parameter $\tilde{t}_s$ is currently unknown. In order to provide a viable alternative to the practitioners, we assume equal misclassification costs and perform model selection using the balanced accuracy score (*BACC*). Table 6 presents the results of the top-performing cost-insensitive models in every prediction period. We note that there is some variation in model performance across periods, which we attribute to changes in the class distribution across different months.

When we compare our results to the literature, we find that a number of studies demonstrate higher performance under more severe class imbalance. For example [11] reports *AUC* 0.76 in a model of predicting sick leave due musculoskeletal disorders at the horizon of 3 months having less than 1% positives in their sample. Their predictors included musculoskeletal complaints, burnout, distress, among others. In [13] measures of depressed mood, distress and fatigue are used, whereas in [12] self-rated health, mental health factors and psychosocial work characteristics are described, and finally, [14] shows the highest performance *AUC* 0.86 (10% positives) using an attitudinal predictor called the Work Ability Index. We conclude that when the misclassification costs are unknown, some objective health-related predictors may be necessary to achieve better performance.

5.2.1. Cost performance of cost-insensitive models

To investigate the cost performance of the top-ranking cost-insensitive models, we calculated cost metrics under all nine intervention scenarios. Fig. 2 demonstrates the difference in cost performance achieved by the top-ranking cost-insensitive models when ranked by *BACC* versus top-ranking cost-sensitive models. Despite reasonable performance, when evaluated using error-based metrics, cost-insensitive models rarely have positive *CIS* and are always inferior to cost-sensitive models.

Model selection based on *BACC* does not appear to be a viable solution once costs are known, based on costs derived using our artificial $\tilde{t}_s$. Instead, effort should be put into determining $\tilde{t}_s$ to facilitate the use

Table 6
Top-model selection when ranking by balanced accuracy. *CIS* for each combination of intervention and $\tilde{t}_s$ are included for reference. Rand refers to record-dependent $p_i \in [0,1]$.

Target	Method		Error-based metrics				Cost-Improvement Score (CIS)								
Period	algo_sampling_calib_T	% pos	FNR	FPR	BACC	AUC	Case1, $\tilde{t}_s=$			Case2, $\tilde{t}_s=$			Case3, $\tilde{t}_s=$		
							1	0.5	rand	1	0.5	rand	1	0.5	rand
2018/03	adab1_none_isotonic_Tbest	15.57	0.42	0.33	0.62	0.67	0.22	0.23	0.15	−3.98	−2.56	−2.38	−0.67	−0.23	−0.18
2018/04	easy_ensmb_none_none_Tbest	8.13	0.40	0.35	0.63	0.67	0.24	0.12	0.06	−1.68	−0.96	−0.77	−0.01	0.21	0.23
2018/05	easy_ensmb_none_none_Tbest	8.0	0.37	0.36	0.63	0.68	0.00	−0.19	−0.84	−3.16	−2.53	−2.33	−0.47	−0.28	−0.23
2018/06	easy_ensmb_none_none_Tbest	9.12	0.37	0.34	0.65	0.70	0.23	0.19	0.16	−3.67	−2.88	−2.62	−0.61	−0.36	−0.35
2018/07	adab1_none_isotonic_Tbest	6.87	0.42	0.31	0.63	0.69	0.04	−0.16	−0.30	−1.87	−1.29	−1.12	−0.05	0.13	0.16
2018/08	adab1_ros_none_Tbest	6.8	0.42	0.27	0.65	0.71	0.11	−0.01	−0.17	−2.73	−2.07	−1.85	−0.28	−0.07	−0.04
2018/09	easy_ensmb_none_none_Tbest	9.26	0.45	0.29	0.63	0.68	0.34	0.45	0.49	−3.14	−2.23	−2.08	−0.40	−0.12	−0.05
2018/10	adab1_ros_isotonic-0.5	11.61	0.42	0.32	0.63	0.68	0.28	0.14	0.10	−3.30	−2.23	−2.06	−0.47	−0.14	−0.09
2018/11	adab1_ros_none-0.5	10.86	0.41	0.32	0.63	0.68	0.15	−0.02	−0.14	−3.87	−2.89	−2.73	−0.64	−0.34	−0.30
2018/12	adab1_none_isotonic_Tbest	9.04	0.35	0.36	0.64	0.69	0.10	−0.15	−0.19	−2.20	−1.45	−1.17	−0.18	0.06	0.10
2019/01	easy_ensmb_none_sigmoid_Tbest	12.98	0.45	0.31	0.62	0.66	0.12	−0.09	−0.23	−3.81	−2.56	−2.35	−0.61	−0.22	−0.13
2019/02	easy_ensmb_none_isotonic_Tbest	14.75	0.39	0.35	0.63	0.67	0.15	0.11	0.04	−4.54	−3.21	−2.74	−0.85	−0.44	−0.34
2019/03	easy_ensmb_none_none_Tbest	11.71	0.39	0.32	0.65	0.70	0.28	0.12	−0.08	−2.18	−1.21	−1.12	−0.13	0.17	0.22

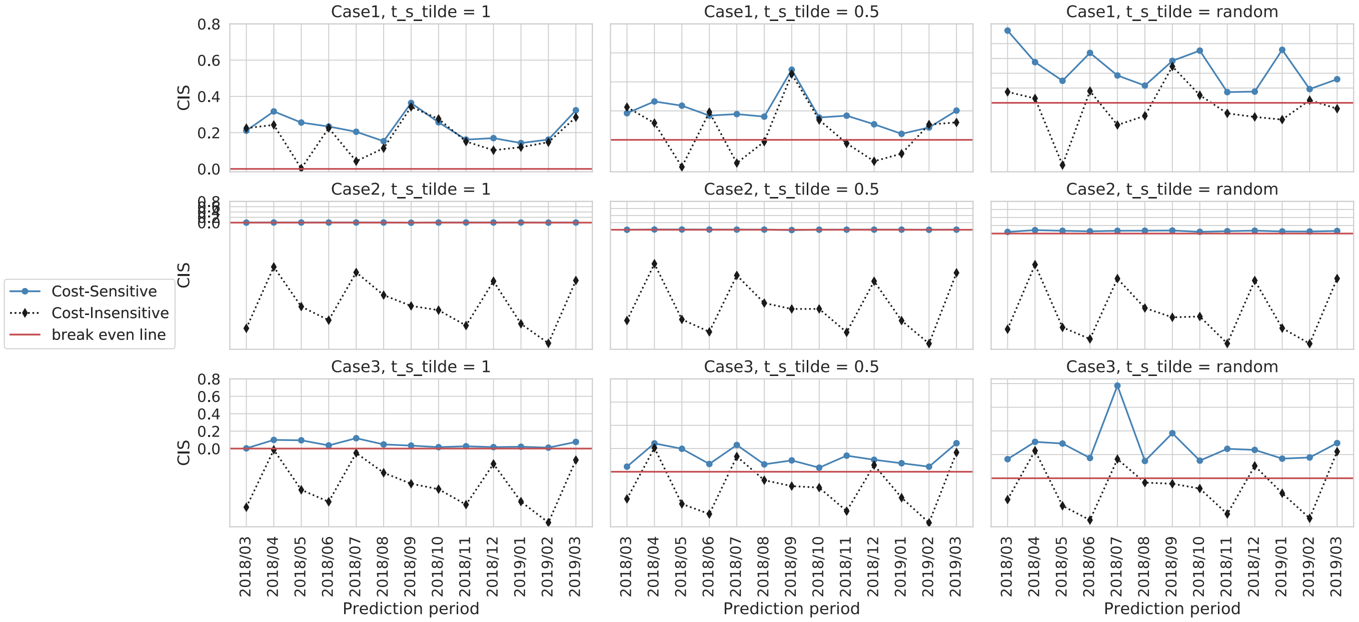

Fig. 2. Comparison of top models' cost performance across prediction periods and interventions. The solid line represents the best performing cost-sensitive models (ranked by *CIS*). The dotted line shows the best performing cost-insensitive model (ranked by *BACC*) under the same operating conditions. The horizontal line shows $CIS = 0$. Here "t_s_tilde" is $\tilde{t}_s/t_s$ and "random" refers to record-dependent $p_i \in [0,1]$.

of misclassification costs. Of course, once this is done, this statement can be revisited.

6. Conclusion

In this paper we describe a potential solution to a real-world problem, requested by HR and Well-being specialists. We were provided with an anonymised dataset containing employee payroll information only, without any health-related data. This carried the risk of the resulting models be based on correlations unrelated to health. We therefore emphasise that our conceptual model was developed under the assumption that any intervention considered aims to *increase employee well-being*. Thus targeting someone erroneously should never lead to a negative outcome for that individual.

Our focus was on developing a flexible, algorithm-agnostic framework to predict short-term employee sickness absence using HR and payroll data only.

Our definition of the target variable, which depends on the threshold T_{hrs} of number of hours of sickness absence can be adapted as needed. In fact, it is possible to use the same decision support system with a range of targets to focus on absences of specific duration (e.g. short term versus long term).

Thanks to relying on existing open-source implementations, our model selection and evaluation framework can be readily deployed to the cloud. We provided two alternative evaluation metrics under class and cost imbalance: with and without knowledge of costs. We demonstrate that improved cost-effectiveness of intervention campaigns can be achieved by focussing classifiers on the specific operating conditions of the underlying training data.

Our main take-away for practitioners is as follows:

- In the absence of misclassification costs, balanced accuracy allows one to ensure good model performance on both classes, instead of only the positives, but cost-based metrics almost always achieve higher savings.
- Models selected based on *BACC* are not a cost-effective alternative to those selected on cost-based metrics. Thus efforts should be directed to quantifying the effect interventions have on individuals so that misclassification costs can be specified.
- When the expected response to an intervention is known, *DMECC* allows to assess the cost-effectiveness of an intervention for each observation resulting in a more profitable intervention campaign.
- Investigating the cost distribution on the available training data can allow one to conduct a preliminary cost and benefit analysis of a given intervention.

6.1. Limitations and future research

The experiments presented in this paper are not without limitations, some of which are subject to data availability.

Firstly, the seasonal changes in class imbalance are accounted for by including, for example, sickness absences from other years as features, as well as creating derivative features from that data (e.g. each individual's average absences in February in all previous years).

Secondly, we assume that the effect of any given intervention does not extend beyond the prediction period. This can be rectified by either considering a longer prediction horizon (e.g. a quarter), or by extending the conceptual model to a dynamic setting. This, combined with the use of panel data to help the model cope with seasonal changes in class imbalance, might be an interesting avenue for future research.

Thirdly, our experiments hinge on the *apriori* knowledge of the individual response to the intervention, which in practice can rarely be obtained. Instead of assigning arbitrary values, this parameter could be estimated, with the help of causal inference models. Causal models estimate individual-level treatment effects from observational data or randomized control trials, an example of algorithmic causal modelling can be found in [58]. We note that in this case, the cost matrix will also depend on the performance of the chosen causal model that predicts $\widetilde{t}_s$. Related to that, ways to improve the estimates of expected hours of absence per individual might also be worth revisiting.

Lastly, we assume that employees with varying hours of sickness will have identical response to any given intervention. Without the knowledge of the true cause of absence, it is not possible to know if the absence is at all preventable. It is not realistic to target, for example, an employee with 10 h of absence caused by a common cold and an employee with 100 h of absence caused by burnout equally. To rectify this, it is highly desirable to collect objective health-related information, such as e.g. the causes of absence. An interesting direction could be to consider multi-class setting, evaluating an array of interventions or to apply cost-sensitive regression such as in [59,60] but with real costs.

Acknowledgements

This research was funded by the Institute for the encouragement of Scientific Research and Innovation of Brussels – Innoviris (Project reference: 2017-Explore-61 AI-WHRM).

The data for this research was kindly provided by Attentia, specialist in HR and Well-being.

Computational resources and services were provided by the Shared ICT Services Centre funded by the Vrije Universiteit Brussel, the Flemish Supercomputer Center (VSC).

We also thank the anonymous reviewers for their useful comments that helped improve this work.

References

[1] OECD, Sickness, Disability and Work: Breaking the Barriers, 2010, https://doi.org/10.1787/9789264088856-en.

[2] The World Bank, GDP (current US$), https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?end=2018 locations=EU-US-CN-OE start=2003, 2021.

[3] O. Publishing, Mental Health and Work Fit Mind, Fit Job: From Evidence to Practice in Mental Health and Work, OECD Publishing, 2015, https://doi.org/10.1787/9789264228283-en.

[4] S.G. Aldana, N.P. Pronk, Health promotion programs, modifiable health risks, and employee absenteeism, J. Occup. Environ. Med. 43 (1) (2001) 36.

[5] T. DeGroot, D.S. Kiker, A meta-analysis of the non-monetary effects of employee health management programs, in: Human Resource Management: Published in Cooperation with the School of Business Administration 42 (1), The University of Michigan and in alliance with the Society of Human Resources Management, 2003, pp. 53–69.

[6] K.M. Parks, L.A. Steelman, Organizational wellness programs: a meta-analysis, J. Occup. Health Psychol. 13 (1) (2008) 58.

[7] D. Pessach, G. Singer, D. Avrahami, H.C. Ben-Gal, E. Shmueli, I. Ben-Gal, Employees recruitment: A prescriptive analytics approach via machine learning and mathematical programming, Decis. Support Syst. 134 (2020) 113290.

[8] A. Tursunbayeva, S. Di Lauro, C. Pagliari, People analytics—a scoping review of conceptual boundaries and value propositions, Int. J. Inf. Manag. 43 (2018) 224–247, https://doi.org/10.1016/j.ejor.2018.06.035.

[9] A. Burdorf, Prevention strategies for sickness absence: sick individuals or sick populations? Scand. J. Work Environ. Health 45 (2) (2019) 101–102.

[10] B. Bosman, R. Roelen, H. Heymans, Prediction models to identify workers at risk of sick leave due to low back pain in dutch industry, Eur. J. Pub. Health 26 (2016), https://doi.org/10.1093/eurpub/ckw174.208.

[11] L.C. Bosman, C.A. Roelen, J.W. Twisk, I. Eekhout, M.W. Heymans, Development of prediction models for sick leave due to musculoskeletal disorders, J. Occup. Rehabil. (2019) 1–8, https://doi.org/10.1007/s10926-018-09825-y.

[12] S.F.A. Duijts, I. Kant, J.A. Landeweerd, G.M.H. Swaen, Prediction of sickness absence: development of a screening instrument, Occup. Environ. Med. 63 (8) (2006) 564–569, https://doi.org/10.1136/oem.2005.024521.

[13] M.F.A. van Hoffen, C.I. Joling, M.W. Heymans, J.W.R. Twisk, C.A.M. Roelen, Mental health symptoms identify workers at risk of long-term sickness absence due to mental disorders: prospective cohort study with 2-year follow-up, BMC Public Health 15 (1) (2015), https://doi.org/10.1186/s12889-015-2580-x.

[14] A. Lundin, O. Leijon, M. Vaez, M. Hallgren, M. Torgén, Predictive validity of the work ability index and its individual items in the general population, Scand. J. Publ. Health 45 (4) (2017) 350–356, https://doi.org/10.1177/1403494817702759.

[15] C.A.M. Roelen, M.W. Heymans, J.W. Twisk, M. Laaksonen, S. Pallesen, N. Magerøy, B.E. Moen, B. Bjorvatn, Health measures in prediction models for high sickness absence: single-item self-rated health versus multi-item sf-12, Eur. J. Publ. Health 25 (4) (2015) 668–672, https://doi.org/10.1093/eurpub/cku192.

[16] Z. Szubert, T. Makowiec-Dabrowska, D. Merecz, W. Sobala, Predictors of short- and long-term sickness absence in female post office workers in poland, Int. J. Occup. Med. Environ. Health 29 (4) (2016) 539–562, https://doi.org/10.13075/ijomeh.1896.00795.

[17] C.A.M. Roelen, C.M. Stapelfeldt, M.W. Heymans, W. van Rhenen, M. Labriola, C. V. Nielsen, U. Bültmann, C. Jensen, Cross-national validation of prognostic models predicting sickness absence and the added value of work environment variables, J. Occup. Rehabil. 25 (2) (2015) 279–287, https://doi.org/10.1007/s10926-014-9536-3.

[18] V.S. Araujo, T.S. Rezende, A.J. Guimarães, V. Araujo, P. de Campos Souza, A hybrid approach of intelligent systems to help predict absenteeism at work in companies, SN Appl. Sci. 1 (6) (2019) 536.

[19] R.P. Ferreira, A. Martiniano, D. Napolitano, E.B.P. Farias, R.J. Sassi, Artificial neural network and their application in the prediction of absenteeism at work, Int. J. Rec. Sci. Res. 9 (1) (2018) 23332–23334, https://doi.org/10.24327/IJRSR.

[20] A. Martiniano, R. Ferreira, R. Sassi, C. Affonso, Application of a neuro fuzzy network in prediction of absenteeism at work, in: 7th Iberian Conference on Information Systems and Technologies (CISTI 2012), IEEE, 2012, pp. 1–4.

[21] Z. Wahid, A. Satter, A. Al Imran, T. Bhuiyan, Predicting absenteeism at work using tree-based learners, in: Proceedings of the 3rd International Conference on Machine Learning and Soft Computing, ACM, 2019, pp. 7–11.

[22] C.R.L. Boot, A. van Drongelen, I. Wolbers, H. Hlobil, A.J. van der Beek, T. Smid, Prediction of long-term and frequent sickness absence using company data, Occup. Med. 67 (3) (2017) 176–181, https://doi.org/10.1093/occmed/kqx014.

[23] G.M. Weiss, Mining with rarity: a unifying framework, ACM Sigkdd Explor. Newslett. 6 (1) (2004) 7–19.

[24] G.M. Weiss, F. Provost, Learning when training data are costly: the effect of class distribution on tree induction, J. Artif. Intell. Res. 19 (2003) 315–354.

[25] A.-H. Homaie-Shandizi, V.P. Nia, M. Gamache, B. Agard, Flight deck crew reserve: from data to forecasting, Eng. Appl. Artif. Intell. 50 (2016) 106–114, https://doi.org/10.1016/j.engappai.2016.01.028.

[26] D.J. Hand, Measuring classifier performance: a coherent alternative to the area under the roc curve, Mach. Learn. 77 (1) (2009) 103–123.

[27] C. Elkan, The foundations of cost-sensitive learning, in: International Joint Conference on Artificial Intelligence vol. 17, Lawrence Erlbaum Associates Ltd, 2001, pp. 973–978.

[28] S. Viaene, G. Dedene, Cost-sensitive learning and decision making revisited, Eur. J. Oper. Res. 166 (1) (2005) 212–220.

[29] B. Zadrozny, C. Elkan, Learning and making decisions when costs and probabilities are both unknown, in: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2001, pp. 204–213.

[30] V.S. Sheng, C.X. Ling, Thresholding for making classifiers cost-sensitive, in: AAAI, 2006.

[31] D.R. Velez, B.C. White, A.A. Motsinger, W.S. Bush, M.D. Ritchie, S.M. Williams, J. H. Moore, A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction, Genet. Epidemiol. 31 (4) (2007) 306–315.

[32] G.M. Weiss, K. McCarthy, B. Zabar, Cost-sensitive learning vs. sampling: which is best for handling unbalanced classes with unequal error costs? Dmin 7 (35–41) (2007) 24.

[33] D.M. Powers, Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation, J. Mach. Learn. Technol. 2 (2021) 37–63.

[34] A. Tharwat, Classification assessment methods, Appl. Comput. Inform. 17 (1) (2020), https://doi.org/10.1016/j.aci.2018.08.003.

[35] A. Luque, A. Carrasco, A. Martín, A.D.L. Heras, The impact of class imbalance in classification performance metrics based on the binary confusion matrix, Pattern Recogn. 91 (2019) 216–231.

[36] K.H. Brodersen, C.S. Ong, K.E. Stephan, J.M. Buhmann, The balanced accuracy and its posterior distribution, in: 2010 20th International Conference on Pattern Recognition, IEEE, 2010, pp. 3121–3124.

[37] M. Sokolova, N. Japkowicz, S. Szpakowicz, Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation, in: Australasian Joint Conference on Artificial Intelligence, Springer, 2006, pp. 1015–1021.

[38] Federal Government of Belgium, Gewaarborgd Loon, https://www.socialsecurity.be/citizen/nl/arbeidsongeschiktheid-ongeval-en-beroepsziekte/arbeidsongeschikt-door-ziekte/gewaarborgd-loon., 2021.

[39] L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and Regression Trees, Wadsworth, 1984.

[40] F.J. Hastie, R. Tibshirani, The Elements of Statistical Learning, Springer-Verlag New York, 2009.

[41] L. Breiman, et al., Statistical modeling: the two cultures (with comments and a rejoinder by the author), Stat. Sci. 16 (3) (2001) 199–231.

[42] G. Petrides, W. Verbeke, Misclassification Cost-Sensitive Ensemble Learning: A Unifying Framework. arXiv:2007.07361, 2020.

[43] L. Breiman, Bagging predictors, Mach. Learn. 24 (1996) 123–140.

[44] T.K. Ho, The random subspace method for constructing decision forests, IEEE Trans. Pattern Anal. Mach. Intell. 20 (1998) 832–844.

[45] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.

[46] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Mach. Learn. 63 (2006) 3–42.

[47] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, J. Comput. Syst. Sci. 55 (1) (1997) 119–139.

[48] C. Chao, A. Liaw, L. Breiman, Using Random Forest to Learn Imbalanced Data, Tech. Rep, University of California, Berkley, Department of Statistics, 2004.

[49] X.-Y. Liu, J. Wu, Z. Cheng Zhou, Exploratory under-sampling for class-imbalance learning, in: Sixth International Conference on Data Mining (ICDM'06), 2006, pp. 965–969.

[50] C. Seiffert, T.M. Khoshgoftaar, J.V. Hulse, A. Napolitano, Rusboost: a hybrid approach to alleviating class imbalance, IEEE Trans. Syst. Man Cybern. Syst. Hum. 40 (2010) 185–197.

[51] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, Smote: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.

[52] B. Zadrozny, C. Elkan, Transforming classifier scores into accurate multiclass probability estimates, in: KDD, 2002.

[53] J. Platt, et al., Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, Adv. Large Margin Class. 10 (3) (1999) 61–74.

[54] A. Niculescu-Mizil, R. Caruana, Predicting good probabilities with supervised learning, in: ICML, 2005.

[55] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[56] G. Lemaître, F. Nogueira, C.K. Aridas, Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning, J. Mach. Learn. Res. 18 (17) (2017) 1–5.

[57] C. Groskopf, Contributors, csvkit, 2016.

[58] S. Wager, S. Athey, Estimation and inference of heterogeneous treatment effects using random forests, J. Am. Stat. Assoc. 113 (523) (2018) 1228–1242.

[59] M. Czajkowski, M. Czerwonka, M. Kretowski, Cost-sensitive global model trees applied to loan charge-off forecasting, Decis. Support. Syst. 74 (2015) 57–66.

[60] H. Zhao, A.P. Sinha, G. Bansal, An extended tuning method for cost-sensitive regression and forecasting, Decis. Support. Syst. 51 (3) (2011) 372–383.

Natalie Lawrance is a Ph.D. candidate at the Vrije Universiteit Brussel. She holds a M.Sc. in Economics from KU Leuven. Her research interests lie at the intersection of the fields of Applied Economics and Machine Learning.

George Petrides is a senior researcher at the University of Bergen, Norway. He received his PhD in Mathematics at the University of Manchester (UK, 2006). Previously, he was a senior researcher at VUB, Belgium, and has lectured at various academic institutions in Cyprus and at NTNU in Norway, where he also was a post-doctoral fellow. His research interests lie within the fields of Machine Learning and Cryptology.

Marie-Anne Guerry, is full professor at the Vrije Universiteit Brussel, Belgium. She obtained her PhD in Sciences in 1992. Currently, she is teaching mathematics at the Faculty of Social Sciences & Solvay Business School of the Vrije Universiteit Brussel. Her research activities are situated in the domain of Manpower Planning. This research covers on the one hand Markov models and their applications in quantitative Human Resources Management, and on the other hand HR-analytics and empirical research on career progress and career characteristics.