Received 5 April 2017
Revised 30 June 2017
Accepted 23 August 2017

Predicting voluntary turnover through human resources database analysis

Evy Rombaut and Marie-Anne Guerry
Vrije Universiteit Brussel, Brussels, Belgium

Abstract

Purpose – This paper aims to question whether the available data in the human resources (HR) system could result in reliable turnover predictions without supplementary survey information.

Design/methodology/approach – A decision tree approach and a logistic regression model for analysing turnover were introduced. The methodology is illustrated on a real-life data set of a Belgian branch of a private company. The model performance is evaluated by the area under the ROC curve (AUC) measure.

Findings – It was concluded that data in the personnel system indeed lead to valuable predictions of turnover.

Practical implications – The presented approach brings determinants of voluntary turnover to the surface. The results yield useful information for HR departments. Where the logistic regression results in a turnover probability at the individual level, the decision tree makes it possible to ascertain employee groups that are at risk for turnover. With the data set-based approach, each company can, immediately, ascertain their own turnover risk.

Originality/value – The study of a data-driven approach for turnover investigation has not been done so far.

Keywords Human resource management, Turnover, Logistic regression, Database analysis, Human resources analytics, Manpower planning, Wastage

Paper type Research paper

1. Introduction

Predicting voluntary turnover and turnover intentions is an important topic of study. Maintaining long-term skilled employees is crucial for every company (Carmeli and Weisberg, 2006). Also, the voluntary turnover of employees can cost the organization over one and a half times the employee's annual salary, if overall costs are taken into account such as reassigning tasks, recruiting and training replacements (Perryer *et al.*, 2010). Next, it has been shown that companies with a high voluntary turnover rate have significantly lower performances than their rivals (Felps *et al.*, 2009). This may easily endanger a company's future opportunities in the marketplace (Sexton *et al.*, 2005).

Employee turnover can be situated in the domain of manpower planning. Manpower planning models, based on estimations of transition probabilities characterizing employee mobility, are used to predict future staff compositions. These predictions are necessary to ensure that the right people will be available at the right positions at the right times (Bartholomew *et al.*, 1991; Khoong, 1996). To this end, both internal and external transitions should be taken into account. There are two types of external transitions: incoming external transitions and outgoing external transitions. The latter can be referred to as wastage.

The authors thank the reviewers for their remarks and valuable suggestions.

Management Research Review
Vol. 41 No. 1, 2018
pp. 96-112

2040-8269
DOI 10.1108/MRR-04-2017-0098

According to Bartholomew *et al.* (1991), manpower wastage is the most fundamental issue in manpower planning owing to the fact that wastage creates vacancies, and therefore provides opportunities for promotion and recruitment (Ugwuowo and McClean, 2000). Successful manpower planning will thus depend to a great extent on describing and predicting wastage within a company (Ugwuowo and McClean, 2000). This paper will focus on predicting external voluntary outgoing transitions, i.e. if the employee decides to leave the company voluntarily. This type of wastage is referred to as voluntary turnover. Voluntary turnover is the most problematic issue for companies, as this decision is out of the company's control (Sexton *et al.*, 2005). To gather insights on voluntary turnover, often, turnover intention is investigated. Turnover intention is "the behavioural tendency of employees to attempt to leave their work organization, which may lead to actual turnover" (Chen *et al.*, 2014).

Research on voluntary turnover and turnover intention is mostly based on survey data (Carmeli and Weisberg, 2006; Chen *et al.*, 2014; Crossley *et al.*, 2007; Felps *et al.*, 2009; Krau, 1981; Sirola, 1998; Mitchell *et al.*, 2001; Nyberg, 2010; Oshagbemi, 2000; Ramesh and Gelfand, 2010; Singh and Schwab, 2000; Sirvanci, 1984). Analysing actual turnover requires longitudinal data to see whether the employee left in a certain time span. This is why, in many cases, turnover intention, and not actual turnover, is investigated. Performing a survey is also very time-consuming and often results in low response rates, which makes it difficult to draw generalized conclusions. Furthermore, a survey is mostly restricted to a certain company (or sector). However, different determinants will come to the surface across various companies (sectors) (Carmeli and Weisberg, 2006).

In the past decade, organizations have been gathering more and more data, which is not being used, thus far, to its full potential. The human resources (HR) departments thus have a wide variety of data at their disposal (Harris *et al.*, 2011). In the study for customer churn, data mining tools are used to extract useful information from the data set (Verbeke *et al.*, 2011). In doing so, organizations are able to answer "business questions that have been traditionally too time-consuming to solve" (Hung *et al.*, 2006). Customer churn is a very similar phenomenon to employee turnover. The present paper advocates a study of a data-driven approach for turnover investigation, which has not been done so far.

In the past decade, there has been a growing interest in HR analytics. This type of research is mostly focussed on predicting the future economic value of companies (Jac, 2010). According to Angrave *et al.* (2016), the development of HR analytics is blocked by "a lack of understanding of analytical thinking by the HR profession". The authors state that a different approach to HR analytics is needed. This all starts with the question of how an existing personnel data set can be used in a meaningful way to inform HR practice.

The current paper aims to develop a methodological approach to investigate and predict turnover by analysing the available data in a HR database, without additional surveys. The goal of the paper is to fully use the available data for predictive purposes. The advantage of the data set-based approach is that it enables the investigation of turnover (not turnover intention), and its determinants, based on longitudinal data. Also, no supplementary data collection is needed and a much larger group of employees is observed. We acknowledge, of course, the importance and need for surveys to investigate the underlying psychological reasoning of employees for turnover, and for understanding certain phenomena. The data-driven approach that is offered in this paper can be used in a complementary way by companies to ascertain risk groups and study turnover within their own organizations.

The present paper is outlined according to a typical data mining framework used for customer churn (Hung *et al.*, 2006). First, raw data in a HR database is evaluated: the variables – present in the data set – that can be used to predict turnover are identified. To

this end, Section 2 consists of an overview of factors that have been found to influence turnover through survey data. Second, the data are extracted and transformed for the chosen model. In Sections 3 and 4, a discussion follows on the choice of a suitable model for the predictive analysis of turnover. Third, the model needs to be evaluated, therefore, Section 5 elaborates on model evaluation. Lastly, the obtained results need to be interpreted. In Section 6, the useful tools will be illustrated and compared through a real-life data set. In Section 7, limitations and further research are discussed.

2. Factors influencing turnover

In this section, the variables in the data set are identified that can be useful to differentiate between "leavers" and "stayers" in the company.

A general HR data set consists of typical demographic factors (gender, age, marital status, etc.) and work-specific factors (seniority, pay, work percentage, etc.). In literature, various demographic and work-specific factors have been shown to influence employee turnover. However, the direct link of these factors to turnover is not always established; often, turnover intention is investigated. Also, certain demographic and work-specific factors effect job satisfaction and organizational commitment, which in turn, influence turnover. These factors, therefore, have an indirect link to turnover.

2.1 Direct link to turnover intention and turnover

2.1.1 Gender. In their meta-analysis, Griffeth *et al.* (2000) found a similar turnover rate for men and women. This appeared to be even more equal for higher educated women. Older women, however, had a lower turnover rate. In a literature review on teacher mobility, Grissom *et al.* (2016) conclude that in past research, women were generally found to have a higher turnover rate than men. However, they found the opposite to be true in more recent research.

2.1.2 Age. Pitts *et al.* (2011) showed that turnover intention first increases and then, at a certain age decreases again. However, in most other research, turnover intention is found to be negatively related to age (for the entire lifespan) (Griffeth *et al.*, 2000; Carmeli and Weisberg, 2006; Ng and Feldman, 2009). Remarkably, for teachers, the relationship between age and turnover was found to be U-shaped (Grissom *et al.*, 2016).

2.1.3 Seniority. Griffeth *et al.* (2000) state that age is often studied in interaction with seniority. They also find that a shorter length of service leads to a higher turnover intention. Job instability during the first years of employment was found to be significant by Singh and Schwab (2000). Furthermore, a low average level of seniority from previous jobs, and thus a high frequency of job-hopping, is a strong indicator that the employee will be unlikely to stay for a long period of time (Singh and Schwab, 2000).

2.1.4 Pay. From the research by Sirola (1998), it became clear that pay satisfaction has a direct effect on turnover intent. However, not only the quantity of compensation is important, but also the pay growth. In their meta-analysis, Griffeth *et al.* (2000) discuss the effect of pay and pay satisfaction. They found the effect size on turnover to be more modest than how it is normally conceived in literature. Employees should most importantly have a feeling of fairness with regards to their pay. A more recent study among Chinese managers by Guan *et al.* (2014) found that the pay was negatively related to the turnover intention.

2.1.5 Marital status. Marital status, number of children and responsibilities that come along with having a family, may also have an influence on turnover, as these family responsibilities create a need for stability (Krau, 1981). This is also confirmed by a meta-analysis from Griffeth *et al.* (2000). Moreover, Griffeth *et al.* (2000) wrote that the interaction between the family responsibility variables and the gender might also have a predictive

value. Indeed, Valcour and Tolbert (2003) state that, among dual-earning couples, women are more likely to turnover than men.

2.1.6 Nationality. According to Williams and O'Reilly (1998), employees belonging to a racial minority in a company will have a higher rate of turnover. Griffeth *et al.* (2000), however, do not reach unambiguous findings after combining the outcomes of different studies. They find no decisive effect of nationality on turnover. This varying effect was confirmed by Grissom *et al.* (2016).

2.2 Indirect link through job satisfaction and organizational commitment

Job satisfaction, together with organizational commitment, are the two most frequently studied influences on turnover (Allen and Meyer, 1996; Mitchell *et al.*, 2001; Meyer *et al.*, 2002). Job satisfaction can be conceived of as a "multidimensional concept that includes a set of favourable and unfavourable feelings by which employees perceive their job" (Garcıa-Bernal *et al.*, 2005). This satisfaction is influenced by personal characteristics of the employee and characteristics of the job itself (Garcıa-Bernal *et al.*, 2005). Organizational commitment consists of affective, normative and continuance commitment to the company (Allen and Meyer, 1996; Ito and Brotheridge, 2005; Perryer *et al.*, 2010). Affective commitment relates to the emotional attachment which is formed by sharing values with the organization and members of the organization (Allen and Meyer, 1996; Perryer *et al.*, 2010), whereas normative commitment is a sense of obligation towards the company. Continuance commitment comes from the employee's perception of having no alternatives to continuing to work for the current company and the recognition of costs related with leaving the organization (Allen and Meyer, 1996; Perryer *et al.*, 2010). Mitchell *et al.* (2001) introduce the term "job embeddedness". This term represents a broad set of influences on employee retention, containing links to and fit in the organization or community and including sacrifices that will need to be made when leaving a job (e.g. giving up benefits, colleagues, etc.). Research among the US federal employees from Pitts *et al.* (2011) also found that workplace satisfaction plays the largest role in predicting turnover intentions.

Next, the demographic and work-specific factors that influence job satisfaction and organizational commitment are discussed, and thus indirectly are affecting voluntary turnover.

2.2.1 Gender. D'Addio *et al.* (2007) performed an analysis on the determinants of job satisfaction. They found that the determinants of reported job satisfaction differ between genders. This difference is widely studied (Bender and Heywood, 2006; Cohrs *et al.*, 2006; D'Addio *et al.*, 2007; Garcıa-Bernal *et al.*, 2005). Women are generally characterized by a higher job satisfaction (Bender and Heywood, 2006). For men, important determinants of job satisfaction are employer-provided training, gross hourly wages, good health (measured by number of nights spent in a hospital) and working full-time instead of part-time (D'Addio *et al.*, 2007). For women, the only significant determinant for job satisfaction found by D'Addio *et al.* (2007) was employment in the public sector. This gender difference was also perceived by Cohrs *et al.* (2006). In a recent cross-national study by Hauret and Williams (2017), in only three of the 14 studied countries, women were found to report a significant higher job satisfaction than men.

2.2.2 Age. Job satisfaction is generally believed to have a linear relationship with age (Kalleberg and Loscocco, 1983). Clark *et al.* (1996), however, found that job satisfaction has a U-shaped relation with age. If other control variables are left out, they noticed a decline until the age of 31, after which point it picks up again. Later, in their meta-analysis, Ng and Feldman (2010) found a positive relationship between age and job satisfaction. They did, however, find a negative relation to satisfaction regarding promotion possibilities. The latter

can be explained because older workers believe that they have fewer promotion opportunities. Nevertheless, older workers are overall found to be more satisfied with their jobs than younger works (Ng and Feldman, 2010; Zacher and Griffin, 2015).

2.2.3 Seniority. In the previous section, it became clear that a low seniority is a strong indicator for turnover. A study from Oshagbemi (2000) links the level of job satisfaction to the length of service. Moreover, important determinants found in this research were autonomy (Skaalvik and Skaalvik, 2014), participatory leadership and qualification possibilities, which usually grow along with seniority. This is enforced by Garcıa-Bernal *et al.*, 2005, who identify "personal development on the job" to be influential. Wright and Bonett (2002) state that employees that just start working in a company are very motivated and try to strengthen the commitment to the organization. However, they might quickly become disillusioned if the job does not turn out as they expected. This suggests a U-shaped relation, analogous to the age – job satisfaction relationship.

2.2.4 Pay. Artz (2008) found that the pay based on individual performances in big firms increases job satisfaction for male employees, as men are more competitive than women. According to Sirvanci (1984), a perceived low salary is one of the main reasons for turnover. Moreover, Nyberg (2010) found that even employees with high job satisfaction will be more likely to leave if they experience a low pay growth.

2.2.5 Nationality. The employee's nationality and culture are important antecedents for job satisfaction (Yousef, 2000; Al-Aameri, 2000). If the company culture is closer to the employee's culture, this will influence job satisfaction and strengthen organizational commitment.

2.2.6 Work percentage. Higgins *et al.* (2000) found that part-time work avoids the work – family conflict for women, and thus increases job satisfaction. Also, both job satisfaction and organizational commitment raises when flexible working hours are offered in a company (Scandura and Lankau, 1997). For men, job satisfaction was found to be higher for full-time workers than for part-time workers (D'Addio *et al.*, 2007).

2.2.7 Sector. The field of employment also proved to be important, as a difference was found in the public and the private sector (D'Addio *et al.*, 2007). It also appeared that qualification possibilities and development were less important in the public sector to achieve job satisfaction (Cohrs *et al.*, 2006).

2.2.8 Workload. Cohrs *et al.* (2006) found another important but less measurable factor to be workload and the stress evoked by it. This is especially the case in certain work areas such as teaching in a school or academic environment (Collie *et al.*, 2012). In a meta-analysis on the job satisfaction for nurses, Zangaro and Soeken (2007) find that job stress correlates most strongly with job satisfaction.

2.2.9 Education level. A factor that influences affective commitment found by Ito and Brotheridge (2005) is career adaptability (i.e. providing opportunities for personal development and promotion). Employees with a higher education level generally have higher career adaptability. This adaptability can reduce turnover and also reinforce normative commitment. It can, however, work the other way, as employees might discover their own market value and perceive options in other companies (Ito and Brotheridge, 2005). Continuance commitment, on the other hand, is reached in economically difficult times, when an employee lacks alternatives (Perryer *et al.*, 2010). Also, managers should give extra attention to higher performers, as they will even resign in economically difficult times (with low employment rates) when pay growth is slowed, as these employees are still confident that they can find a better job (Nyberg, 2010).

2.2.10 Health. In their meta-analysis, Faragher *et al.* (2005) combine the outcomes of almost 500 studies on the relationship between job satisfaction and health. The causal

relationship between these two cannot always be inferred. However, the authors do find a strong correlation between the job satisfaction and the mental and physical health of the employee.

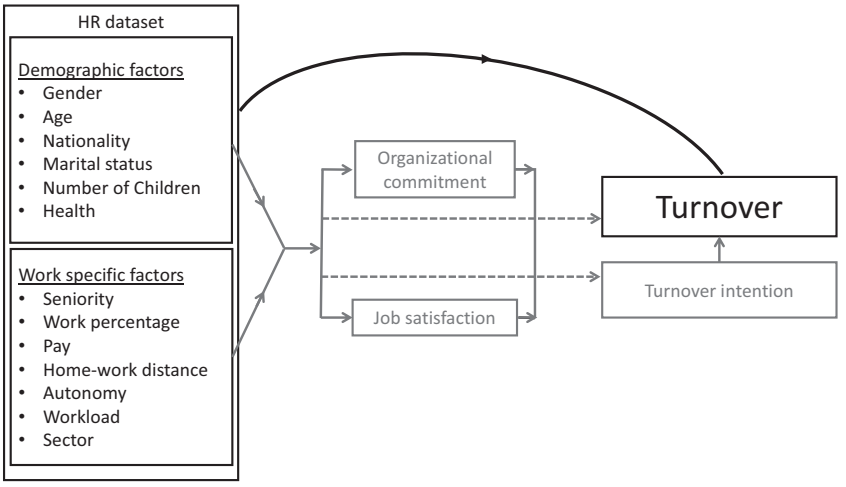
2.2.11 Home–work distance. The duration of the commute (Sirgy *et al.*, 2008), as a result of the work–home distance, is negatively related to job satisfaction. Job satisfaction will be improved with less commuting time, where this can be the cause of living closer by or having the possibility to work from home. A meta-analysis by MacDonald (1999) points out that especially women face commuting issues, caused by their household responsibilities.

It can be concluded that job satisfaction and organizational commitment cannot be measured directly from a HR database. However, they are influenced by a great deal of factors that can be measured and which are present in the database, as is illustrated in Figure 1. The variables in the data set will have a direct and/or indirect influence to employee turnover.

3. Predictive analysis

The previous section discussed factors that influence turnover. These factors will be critical in predicting whether an employee will decide to leave. To be able to choose a technique that leads to reliable predictions, first, techniques that have already been applied in turnover studies will be discussed. Next, a discussion follows on techniques that have been used for customer churn prediction to assimilate the approaches to turnover analysis.

Wang *et al.* (2011) state that HR information management should be strengthened by using inside enterprise (i.e. company- and employee-specific information) and outside enterprise information (i.e. economic circumstances) about the employees' situation. A conceptual framework is given for a decision support system based on turnover knowledge. They provide a very useful evaluation model for employee turnover risk, where both inside and outside factors are taken into account. Because the authors do not elaborate on the



Note: Direct relationship to turnover as studied in the present paper (black)

Figure 1.
Indirect relationship to turnover as studied in literature (grey)

methodology and the application of the framework, they encourage further study of suitable mechanisms. In this paper, the theoretical framework is extended by providing a methodology.

Sexton *et al.* (2005) focus on a neural network approach to predict turnover. They use a neural network algorithm which has been found to be a successful prediction tool for a variety of fields. The algorithm makes it possible to see which inputs in the data set are actual predictors for turnover by assigning weights to each variable. If the weight is set to zero by the algorithm, the variable is said to be non-relevant for the outcome variable. However, because a random seed is set to start the algorithm, it may yield different results when run multiple times (Sexton *et al.*, 2005). Sexton thus let the algorithm run 10 times to determine which variables are important. In the end, only two variables were found to have predictive value: salary and work percentage (full-time vs part-time). So, the question rises as to whether small effects from other variables are neglected by this algorithm. This question remains unanswered, as neural networks are so called "black boxes": it is unclear what happens inside the algorithm. This is often viewed as a disadvantage. Also, from the previous section, it became clear that many other factors can have an influence on turnover, which leads to the investigation of other predictive methods.

When it comes to managing customer churns, companies are interested in predicting which customers are more likely to churn and which incentives they should use to influence customers to stay (Neslin *et al.*, 2006; Verbeke *et al.*, 2011). Neslin *et al.* (2006) studied which methodological approaches work best for predicting customer churn. It appears that logistic regression and decision tree are the most common estimation techniques. Verbeke *et al.* (2011) provide an overview of the literature on churn prediction modelling. From this overview, it is also clear that logistic regression and decision tree are one of the most frequently used techniques for churn prediction. Other distribution functions might be proposed for the analysis of a dichotomous outcome variable other than logistic distribution. Important reasons to choose a logistic distribution are: that it is a flexible and easily used function (Hosmer and Lemeshow, 2004), the model parameters make meaningful estimates possible and it is a well-known technique used in many marketing applications (Neslin *et al.*, 2006).

Therefore, in this paper, a logistic regression model is applied for turnover analysis.

4. Turnover predictive analysis techniques

The goal of this paper is to show that voluntary turnover can be predicted using *a priori* only data available in a HR database and no supplementary information from surveys. Table I shows an example of an extract from a HR database consisting of data from several years. Each year consists of information characterizing every employee in the system and whether or not the employee voluntarily left the company. This is called a case in the data set.

Table I.
Extract example from a HR database

Year	ID	Sex	Marital status	Age	Seniority	Grade	Work(%)	Left
2012	135	Male	Single	41	15	2	100	No
2013	135	Male	Single	42	16	2	100	Yes
2012	136	Female	Divorced	35	9	1	80	No
2013	136	Female	Divorced	36	10	1	80	No
2014	136	Female	Divorced	37	11	1	80	No

In what follows, it is specified in which way logistic regression can be used on a HR database (similar to the example in Table I) for analysing turnover.

4.1 Logistic regression

Regression methods have become an integral component of any data analysis concerned with describing the relationship between a response variable and one or more explanatory variables (Hosmer and Lemeshow, 2004). Therefore, regression analysis can be a key in determining which characteristics have an influence on the transition behaviour of employees in a company. Logistic regression analysis can be used to evaluate the relationship between the observable factors and the turnover probability. If $\hat{Y}$ is the probability of a "yes" for a given set of predictor variables $X_1, X_2, \ldots, X_p$, the logistic regression model prescribes that (Hosmer and Lemeshow, 2004)

$$\hat{Y} = \frac{e^{\beta_0 + \sum_{i=1}^{p} \beta_i X_i}}{1 + e^{\beta_0 + \sum_{i=1}^{p} \beta_i X_i}}$$

In the context of turnover analysis, the dichotomous variable Y will result in the probability of voluntary leaving $\hat{Y}$.

5. Model evaluation

The performance of the logistic regression model will be evaluated with the AUC-measure. This measure is "especially useful for domains with skewed class distribution" (Fawcett, 2006). Because the data set is unbalanced with regards to voluntary turnover, the estimated turnover probabilities are generally low. The idea of the predictive model is that HR managers can target employees with retention strategies who have a turnover probability above a certain threshold (e.g. 30 per cent), which is considered "too high".

The logistic regression leads to individual turnover probabilities for each of the employees. This means that a function m: $X \rightarrow [0,1]$ is obtained which maps a case in the data set to the probability of leaving the company. A case $x \in X$ is an array of factors that represent an employee in the company at a given time. However, to evaluate the model, the resulting probabilities should be compared to the binary outcome 0 or 1, which indicates whether the employee actually left or not. To do so, a classifier is needed that maps a case to 0 or 1 depending on the estimated probability. A model can be converted to a classifier by choosing a certain threshold value t. Given an estimated probability $s = m(x)$, the case is classified in class 1 if $s > t$, and in class 0 otherwise (Hernandez-Orallo *et al.*, 2012). The choice of the threshold will lead to a discrete classifier that maps a case to 0 or 1 depending on the estimated probability.

Instead of 0 and 1, classes are often labelled with P and N (Fawcett, 2006), which respectively stand for positive and negative. With these labels, the possible outcomes of the classifier can be easily described. There are four possible outcomes of a classifier. If the case is positive and is classified as positive, it is said to be a true positive; if it is classified as negative, it is said to be a false negative. If the case is negative and is classified as a negative, it is called a true negative; if it is classified as a positive, it is called a false positive (Table II).

The performance of the classifier can then be evaluated with several measures. The true positive (tp) rate or sensitivity of the classifier is the proportion of positives that have been correctly classified. The false positive (fp) rate of the classifier is the proportion of negatives that have been classified as positives (Fawcett, 2006; Hernandez-Orallo *et al.*, 2012). Another common term is specificity which is defined as 1 − fp rate. For a certain threshold, the rates

tp and fp can be presented as a couple in the two-dimensional ROC (receiver operating characteristics) space. In ROC space, the fp rate is plotted on the x-axis and the tp rate on the y-axis. Calculating the rates for each possible threshold $t \in [0,1]$ leads to a ROC curve. The diagonal line $y = x$ represents a classifier that would randomly guess a class (Fawcett, 2006). The ROC curve of a good classifier should thus lie above this line.

The ROC performance can also be reduced to a single scalar value by calculating the area under the ROC curve (AUC) (Fawcett, 2006). A classifier that performs better than random guessing thus has an AUC-value above 0.5. A perfect classifier will have 1 as the AUC-value. The AUC-value can be used to evaluate the performance of a classifier, but can, however, not be used for direct model comparison (Ferri *et al.*, 2011; Hand and Anagnostopoulos, 2013).

6. Illustration

The illustration is made on a data set from a Belgian branch of a private company that specializes in HR solutions. The turnover probabilities for their own employees are investigated. The data set consists of 13,484 cases, representing 4,041 individuals over the time frame 2006 to 2016, with the following variables:

- Demographic information: gender, age, seniority, marital status, number of children, education and nationality. The variable marital status is recoded to a dichotomous variable partner (yes or no). The variable education is used as a dichotomous variable where 0 = "secondary school" and 1 = "higher education" (BA/MA/PhD). Lastly, nationality is also used as a dichotomous variable where 0 = "Belgian" and 1 = "other".
- Work-specific information: salary, company car (yes or no), company phone (yes or no), internet at home (yes or no) (which are all measures for pay), number of sick days in that year (measure for health) and work regime. The value work regime is an integer ranging from 4 to 40, being the number of hours an employee has to work according to his/her contract.

A logistic regression model was built through forward backward selection of the variables (Hosmer and Lemeshow, 2004). The time frame of the obtained data is 2006 to 2016. For the analysis, the data are pooled over these years. This is done for multiple reasons. The first reason is to be able to fully use all the obtained information. If the regression is built on 1 year only, the number of voluntary leavers is not high enough to find a pattern within the leavers, because a data set will generally be severely unbalanced with regards to the stayer–leaver rate. Second, the goal of this paper is to find the objective factors that lead to voluntary turnover over the course of several years – factors that are not dependent on a certain economic situation in any particular year. However, environment and time changes can have an influence on turnover decisions as well. Therefore, to check for possible environment changes in the pooled data set, a time dummy was included, once with reference year 2006 and reference year 2016. From this analysis, it appeared that two year

Table II. Possible outcomes for a classifier

	True class	
	P	N
Predicted class		
P	True positive	False positive
N	False negative	True negative

groups can be formed in this data set. In the first year group (2006, 2010, 2012, 2013 and 2014), there was a significant lower turnover rate than in the other year group (2007, 2008, 2009, 2011, 2015 and 2016). However, building the regression model for both year groups separately resulted in the following findings:

- the majority of variables remain significant; and
- the signs of the coefficients are the same as in the pooled regression.

Therefore, in the regression model below, all years are pooled.

A variety of variables from the data set are highly significant for the prediction of turnover as can be seen in Table III.

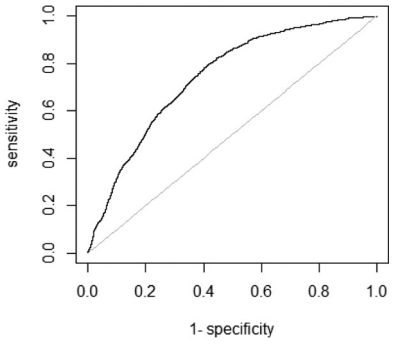
The ROC curve of the regression method lies above the diagonal line, as can be seen in Figure 2. The AUC is 0.7432, so the data set does indeed hold characteristics with valuable predictive information for turnover.

In what follows, the relations between the dependent and independent variables are addressed. First, the independent variables are discussed, which have a similar relationship to the dependent variable as expected in literature. According to the model, women have a lower turnover probability. In literature, it has indeed been indicated that gender has an

Variable	Estimate	Standard error	*Z* value	*Pr*(>|*z*|)	
(Intercept)	−2.021	3.297e-01	−6.129	8.85e-10	***
SexWoman	−0.3664	5.782e-02	−6.336	2.36e-10	***
Age	0.0112	5.342e-03	2.096	0.0361	*
Seniority	0.3180	4.893e-02	6.500	8.04e-11	***
PartnerYes	−0.3075	5.623e-02	−5.469	4.53e-08	***
NationOther	0.4862	9.645e-02	5.041	4.64e-07	***
Salary	−0.00026	6.382e-05	−4.151	3.31e-05	***
Regime	0.03827	6.870e-03	5.571	2.53e-08	***
CarYes	0.8544	1.703e-01	5.018	5.23e-07	***
PhoneYes	0.2934	6.738e-02	4.355	1.33e-05	***
Seniority:Age	−0.01006	1.338e-03	−7.521	5.42e-14	***
CarYes:Salary	0.00029	7.271e-05	4.009	6.10e-05	***

Notes: Significance codes: ***0.0001; *0.01

Table III.
Regression results



Note: AUC = 0.7432

Figure 2.
ROC curve of the logistic regression method

effect on job satisfaction (Bender and Heywood, 2006; Garcıa-Bernal *et al.*, 2005), in that women are usually more satisfied. This higher satisfaction results in a lower turnover intention and thus a lower turnover rate. Next, increasing seniority and age, as found in literature, causes a decrease in turnover probability (Sirvanci, 1984; Singh and Schwab, 2000; Carmeli and Weisberg, 2006; Pitts *et al.*, 2011). From the regression results it is clear that at first, turnover probability increases with increasing seniority until the employee has a length of service of about five years, at which point it decreases. The current data set thus leads to similar results found by Clark *et al.* (1996). This is illustrated in Figure 3.

Employees without a partner have a higher turnover probability than people with a partner. This is an expected result, as employees without a partner do not have the responsibility to provide stability for their partner (Singh and Schwab, 2000). In literature, it is stated, however, that women among dual-earning couples are more likely to turnover (Valcour and Tolbert, 2003); in the current data set, this interaction effect was not significant.

Employees with a lower work regime have a lower turnover probability than employees with a higher work regime. This can be explained through the fact that employees in a lower work regime usually have a better work–life balance and a lower stress level (Higgins *et al.*, 2000). For work regime, the interaction effect with gender was investigated as well, as full-time working men were found to be more satisfied (D'Addio *et al.*, 2007). However, the analysis did not confirm a significant result for the interaction effect.

The effect of nationality on the dependent variable is also in line with literature. Non-Belgian employees have a higher turnover probability. This can be owing to a lower organizational commitment (Yousef, 2000) and cultural differences (Williams and O'Reilly, 1998; Al-Aameri, 2000).

An unexpected result of the regression analysis is the effect of the company car. A higher salary and higher pay satisfaction result in a lower turnover probability according to literature (Griffeth *et al.*, 2000; Artz, 2008). In the data set of the studied company, higher salary does indeed lead to lower turnover, whereas a company car does not.

However, an interaction effect between car and salary was found significant. In the case of higher salaries: out of the two similar employees with the same salary, the employee with the company car will have a higher turnover probability than the employee without a company car. However, employees with lower salaries and a company car do have a lower turnover than employees with the same salary and no car (Figure 4, the cut-off point lies around €2800 monthly gross salary). This unexpected result can be partially explained by the HR department of the studied company. Employees with a company car and a higher

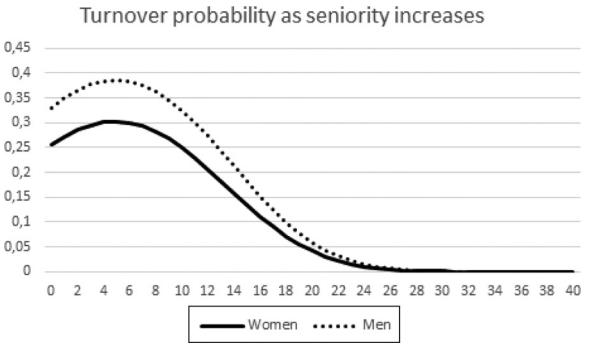


Figure 3.
Turnover probability in relation to the seniority for men and women

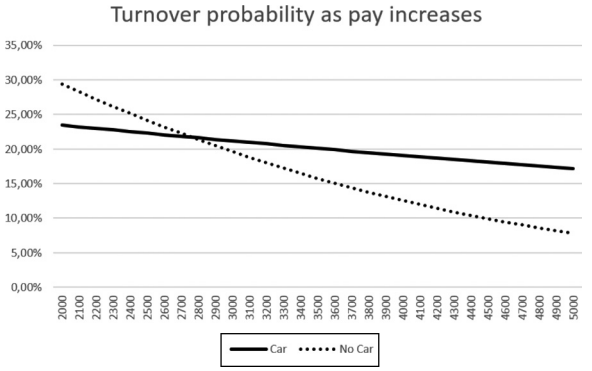


Figure 4.
Evolution of turnover probability with increasing pay

salary are high-performing employees with a high market value. Employees with a high market value can often find better job options or receive better job offers (Ito and Brotheridge, 2005; Nyberg, 2010).

This section is concluded with a Table IV that gives an overview of the variables that were found (not) significant in the data set for predicting turnover.

7. Theoretical implications

It can be concluded that the available data in a company are indeed resourceful for predicting turnover. An important advantage of a data set-based approach is the fact that determinants for actual turnover and not for turnover intention are investigated. Most surveys are performed on employees who currently work in a company and thus give information at the moment of the conducted survey. A longitudinal study is in this case necessary to see which employees leave the company after the survey has been conducted.

Research regarding turnover determinants can now be more developed by combining this data-driven approach with existing methodologies.

8. Practical implications

Notwithstanding the fact that each case of the data set represents a person and consequently, can never be captured entirely by pure raw data, the HR department can use this methodology as a barometer. Departmental managers can use the results of the database analysis to set-up interviews with certain risk groups to prevent turnover. Using the logistic regression obtained from the database, the probability of voluntary turnover for

Significant	Not significant
Gender	Number of children
Age	Education
Seniority	Internet at home
Partner	Number of sick days
Nationality	
Salary	
Work percentage	
Company car	
Company phone	

Table IV.
Overview of significant and not significant variables in the data set

each individual employee can be calculated. Next, HR managers can determine a threshold for the probabilities above which they want to target employees with retention strategies. The information can moreover be used for recruitment purposes. Also, certain company measures can be re-evaluated (e.g. the company car in the current data set that seems to have a reverse effect than would be intuitively suspected), and further investigated.

A recommendation for companies is to expand the database as much as possible and include extra variables such as provided training, an indicator of job autonomy, home–work distance, workload, etc.

9. Limitations

The proposed methodology can, of course, only be used in organizations with a reliable database. Some companies, especially smaller ones, do not have such data. Furthermore, the database might be incomplete or inaccurate. In the prediction of voluntary turnover, it is also of vital importance that the company explicitly indicates in the database whether the employee leaves voluntarily or involuntarily.

As previously stated, a department manager cannot make decisions solely based on the outcomes of the analysis. Similar to every quantitative HR study, it is of vital importance to consider the people behind the data as well.

In the current approach, we did not use information outside the company concerning the economic situation and other job possibilities. This is something that can be taken into account for turnover prediction as well.

10. Further research

With the discussed techniques and the availability of a reliable HR database, a computer package can be developed where the determinants for turnover can be kept up to date and thus, can be used by HR managers to more reliably prevent and reduce voluntary turnover.

For big data scientists and HR analytics, further research could be to include more big data variables such as email traffic (e.g. to other colleagues) or weekly performed hours, which can also prove to be important predictors for voluntary turnover (Morrison, 2004).

When examining turnover, one might also consider looking at a similar phenomenon like survival. In turnover analysis, the probability of leaving is estimated, whereas in survival analysis, the probability of surviving is estimated (in this context, surviving can be interpreted as staying in the company). Survival analysis has been widely used in medical research for predicting survival chances after, for example, a treatment or surgery (Hosmer and Lemeshow, 1999). Also, survival analysis has been linked to wastage in the past by Bartholomew *et al.* (1991). They state that the suitability of the Cox proportional hazards model ought to be investigated empirically in wastage analysis.

References

Al-Aameri, A.S. (2000), "Job satisfaction and organizational commitment for nurses", *Saudi Medical Journal*, Vol. 21 No. 6, pp. 531-535.

Allen, N.J. and Meyer, J.P. (1996), "Affective, continuance, and normative commitment to the organization: an examination of construct validity", *Journal of Vocational Behavior*, Vol. 49 No. 3, pp. 252-276.

Angrave, D., Charlwood, A., Kirkpatrick, I., Lawrence, M. and Stuart, M. (2016), "HR and analytics: why HR is set to fail the big data challenge", *Human Resource Management Journal*, Vol. 26 No. 1, pp. 1-11.

Artz, B. (2008), "The role of firm size and performance pay in determining employee job satisfaction brief: firm size, performance pay, and job satisfaction", *Labour*, Vol. 22 No. 2, pp. 315-343.

Bartholomew, D.J., Forbes, A.F. and McClean, S.I. (1991), *Statistical Techniques for Manpower Planning*, Wiley, New York, NY.

Bender, K.A. and Heywood, J.S. (2006), "Job satisfaction of the highly educated: the role of gender, academic tenure, and earnings", *Scottish Journal of Political Economy*, Vol. 53 No. 2, pp. 253-279.

Carmeli, A. and Weisberg, J. (2006), "Exploring turnover intentions among three professional groups of employees", *Human Resource Development International*, Vol. 9 No. 2, pp. 191-206.

Chen, M.-L., Su, Z.-Y., Lo, C.-L., Chiu, C.-H., Hu, Y.-H. and Shieh, T.-Y. (2014), "An empirical study on the factors influencing the turnover intention of dentists in hospitals in Taiwan", *Journal of Dental Sciences*, Vol. 9 No. 4, pp. 332-344.

Clark, A., Oswald, A. and Warr, P. (1996), "Is job satisfaction U-shaped in age?", *Journal of Occupational and Organizational Psychology*, Vol. 69 No. 1, pp. 57-81.

Cohrs, J.C., Abele, A.E. and Dette, D.E. (2006), "Integrating situational and dispositional determinants of job satisfaction: findings from three samples of professionals", *The Journal of Psychology*, Vol. 140 No. 4, pp. 363-395.

Collie, R.J., Shapka, J.D. and Perry, N.E. (2012), "School climate and social–emotional learning: predicting teacher stress, job satisfaction, and teaching efficacy", *Journal of Educational Psychology*, Vol. 104 No. 4, p. 1189.

Crossley, C.D., Bennett, R.J., Jex, S.M. and Burnfield, J.L. (2007), "Development of a global measure of job embeddedness and integration into a traditional model of voluntary turnover", *Journal of Applied Psychology*, Vol. 92 No. 4, pp. 1031-1042.

D'Addio, A.C., Eriksson, T. and Frijters, P. (2007), "An analysis of the determinants of job satisfaction when individuals baseline satisfaction levels may differ", *Applied Economics*, Vol. 39 No. 19, pp. 2413-2423.

Faragher, E.B., Cass, M. and Cooper, C.L. (2005), "The relationship between job satisfaction and health: a meta-analysis", *Occupational and Environmental Medicine*, Vol. 62 No. 2, pp. 105-112.

Fawcett, T. (2006), "An introduction to ROC analysis", *Pattern Recognition Letters*, Vol. 27 No. 8, pp. 861-874.

Felps, W., Mitchell, T.R., Hekman, D.R., Lee, T.W., Holtom, B.C. and Harman, W.S. (2009), "Turnover contagion: how coworkers' job embeddedness and job search behaviors influence quitting", *Academy of Management Journal*, Vol. 52 No. 3, pp. 545-561.

Ferri, C., Hernandez-Orallo, J. and Flach, P.A. (2011), "A coherent interpretation of AUC as a measure of aggregated classification performance", *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 657-664.

García-Bernal, J., Gargallo-Castel, A., Marzo-Navarro, M. and Rivera-Torres, P. (2005), "Job satisfaction: empirical evidence of gender differences", *Women in Management Review*, Vol. 20 No. 4, pp. 279-288.

Griffeth, R.W., Hom, P.W. and Gaertner, S. (2000), "A meta-analysis of antecedents and correlates of employee turnover: update, moderator tests, and research implications for the next millennium", *Journal of Management*, Vol. 26 No. 3, pp. 463-488.

Grissom, J.A., Viano, S.L. and Selin, J.L. (2016), "Understanding employee turnover in the public sector: insights from research on teacher mobility", *Public Administration Review*, Vol. 76 No. 2, pp. 241-251.

Guan, Y., Wen, Y., Chen, S.X., Liu, H., Si, W., Liu, Y., Wang, Y., Fu, R., Zhang, Y. and Dong, Z. (2014), "When do salary and job level predict career satisfaction and turnover intention among Chinese managers? The role of perceived organizational career management and

career anchor", *European Journal of Work and Organizational Psychology*, Vol. 23 No. 4, pp. 596-607.

Hand, D.J. and Anagnostopoulos, C. (2013), "When is the area under the receiver operating characteristic curve an appropriate measure of classifier performance?", *Pattern Recognition Letters*, Vol. 34 No. 5, pp. 492-495.

Harris, J.G., Craig, E. and Light, D.A. (2011), "Talent and analytics: new approaches, higher ROI", *Journal of Business Strategy*, Vol. 32 No. 6, pp. 4-13.

Hauret, L. and Williams, D.R. (2017), "Cross-national analysis of gender differences in job satisfaction", *Industrial Relations: A Journal of Economy and Society*, Vol. 56 No. 2, pp. 203-235.

Hernandez-Orallo, J., Flach, P. and Ferri, C. (2012), "A unified view of performance metrics: translating threshold choice into expected classification loss", *The Journal of Machine Learning Research*, Vol. 13 No. 1, pp. 2813-2869.

Higgins, C., Duxbury, L. and Johnson, K.L. (2000), "Part-time work for women: does it really help balance work and family?", *Human Resource Management*, Vol. 39 No. 1, pp. 17-32.

Hosmer, D.W. and Lemeshow, S. (1999), *Applied Survival Analysis: Time-to-Event*, Vol. 317, Wiley-Interscience.

Hosmer, D.W. and Lemeshow, S. (2004), *Applied Logistic Regression*, John Wiley & Sons.

Hung, S.-Y., Yen, D.C. and Wang, H.-Y. (2006), "Applying data mining to telecom churn management", *Expert Systems with Applications*, Vol. 31 No. 3, pp. 515-524.

Ito, J.K. and Brotheridge, C.M. (2005), "Does supporting employees' career adaptability lead to commitment, turnover, or both?", *Human Resource Management*, Vol. 44 No. 1, pp. 5-19.

Jac, F.-E. (2010), *The New HR Analytics: Predicting The Economic Value Of Your Company's Human Capital Investments*, AMACOM Div American Mgmt Assn.

Kalleberg, A.L. and Loscocco, K.A. (1983), "Aging, values, and rewards: explaining age differences in job satisfaction", *American Sociological Review*, Vol. 48 No. 1, pp. 78-90.

Khoong, C. (1996), "An integrated system framework and analysis methodology for manpower planning", *International Journal of Manpower*, Vol. 17 No. 1, pp. 26-46.

Krau, E. (1981), "Turnover analysis and prediction from a career developmental point of view", *Personnel Psychology*, Vol. 34 No. 4, pp. 771-790.

MacDonald, H.I. (1999), "Women's employment and commuting: explaining the links", *Cpl Bibliography*, Vol. 13 No. 3, pp. 267-283.

Meyer, J.P., Stanley, D.J., Herscovitch, L. and Topolnytsky, L. (2002), "Affective, continuance, and normative commitment to the organization: a meta-analysis of antecedents, correlates, and consequences", *Journal of Vocational Behavior*, Vol. 61 No. 1, pp. 20-52.

Mitchell, T.R., Holtom, B.C., Lee, T.W., Sablynski, C.J. and Erez, M. (2001), "Why people stay: using job embeddedness to predict voluntary turnover", *Academy of Management Journal*, Vol. 44 No. 6, pp. 1102-1121.

Morrison, R. (2004), "Informal relationships in the workplace: associations with job satisfaction, organisational commitment and turnover intentions", *New Zealand Journal of Psychology*, Vol. 33 No. 3.

Neslin, S.A., Gupta, S., Kamakura, W., Lu, J. and Mason, C.H. (2006), "Defection detection: measuring and understanding the predictive accuracy of customer churn models", *Journal of Marketing Research*, Vol. 43 No. 2, pp. 204-211.

Ng, T.W. and Feldman, D.C. (2009), "Re-examining the relationship between age and voluntary turnover", *Journal of Vocational Behavior*, Vol. 74 No. 3, pp. 283-294.

Ng, T.W. and Feldman, D.C. (2010), "The relationships of age with job attitudes: a meta-analysis", *Personnel Psychology*, Vol. 63 No. 3, pp. 677-718.

Nyberg, A. (2010), "Retaining your high performers: moderators of the performance–job satisfaction–voluntary turnover relationship", *The Journal of Applied Psychology*, Vol. 95 No. 3, pp. 440-453.

Oshagbemi, T. (2000), "Is length of service related to the level of job satisfaction?", *International Journal of Social Economics*, Vol. 27 No. 3, pp. 213-226.

Perryer, C., Jordan, C., Firns, I. and Travaglione, A. (2010), "Predicting turnover intentions: the interactive effects of organizational commitment and perceived organizational support", *Management Research Review*, Vol. 33 No. 9, pp. 911-923.

Pitts, D., Marvel, J. and Fernandez, S. (2011), "So hard to say goodbye? Turnover intention among us federal employees", *Public Administration Review*, Vol. 71 No. 5, pp. 751-760.

Ramesh, A. and Gelfand, M.J. (2010), "Will they stay or will they go? The role of job embeddedness in predicting turnover in individualistic and collectivistic cultures", *The Journal of Applied Psychology*, Vol. 95 No. 5, pp. 807.

Scandura, T.A. and Lankau, M.J. (1997), "Relationships of gender, family responsibility and flexible work hours to organizational commitment and job satisfaction", *Journal of Organizational Behavior*, Vol. 18 No. 4, pp. 377-391.

Sexton, R.S., McMurtrey, S., Michalopoulos, J.O. and Smith, A.M. (2005), "Employee turnover: a neural network solution", *Computers & Operations Research*, Vol. 32 No. 10, pp. 2635-2651.

Singh, D.A. and Schwab, R.C. (2000), "Predicting turnover and retention in nursing home administrators' management and policy implications", *The Gerontologist*, Vol. 40 No. 3, pp. 310-319.

Sirgy, M.J., Reilly, N.P., Wu, J. and Efraty, D. (2008), "A work-life identity model of well-being: towards a research agenda linking quality-of-work-life (QWL) programs with quality of life (QOL)", *Applied Research in Quality of Life*, Vol. 3 No. 3, pp. 181-202.

Sirola, W. (1998), "Explaining nursing turnover intent: job satisfaction, pay satisfaction, or organizational commitment?", *Journal of Organizational Behavior*, Vol. 19, pp. 305-320.

Şirvanci, M. (1984), "Forecasting manpower losses by the use of renewal models", *European Journal of Operational Research*, Vol. 16 No. 1, pp. 13-18.

Skaalvik, E.M. and Skaalvik, S. (2014), "Teacher self-efficacy and perceived autonomy: relations with teacher engagement, job satisfaction, and emotional exhaustion", *Psychological Reports*, Vol. 114 No. 1, pp. 68-77.

Ugwuowo, F. and McClean, S. (2000), "Modelling heterogeneity in a manpower system: a review", *Applied Stochastic Models in Business and Industry*, Vol. 16 No. 2, pp. 99-110.

Valcour, P.M. and Tolbert, P. (2003), "Gender, family and career in the era of boundarylessness: determinants and effects of intra- and inter-organizational mobility", *International Journal of Human Resource Management*, Vol. 14 No. 5, pp. 768-787.

Verbeke, W., Martens, D., Mues, C. and Baesens, B. (2011), "Building comprehensible customer churn prediction models with advanced rule induction techniques", *Expert Systems with Applications*, Vol. 38 No. 3, pp. 2354-2364.

Wang, X., Wang, H., Wang, H., Zhang, L. and Cao, X. (2011), "Constructing a decision support system for management of employee turnover risk", *Information Technology and Management*, Vol. 12 No. 2, pp. 187-196.

Williams, K.Y. and O'Reilly, C.A. III. (1998), "A review of 40 years of research", *Research in Organizational Behavior*, Vol. 20, pp. 77-140.

Wright, T.A. and Bonett, D.G. (2002), "The moderating effects of employee tenure on the relation between organizational commitment and job performance: a meta-analysis", *Journal of Applied Psychology*, Vol. 87 No. 6, pp. 1183.

Yousef, D.A. (2000), "Organizational commitment: a mediator of the relationships of leadership behavior with job satisfaction and performance in a non-Western country", *Journal of Managerial Psychology*, Vol. 15 No. 1, pp. 6-24.

Zacher, H. and Griffin, B. (2015), "Older workers age as a moderator of the relationship between career adaptability and job satisfaction", *Work, Aging and Retirement*, Vol. 1 No. 2, pp. 227-236.

Zangaro, G.A. and Soeken, K.L. (2007), "A meta-analysis of studies of nurses' job satisfaction", *Research in Nursing & Health*, Vol. 30 No. 4, pp. 445-458.

Corresponding author
Evy Rombaut can be contacted at: erombaut@vub.ac.be