Prediction of job transition using publicly available professional profiles

Evie Izeboud
ANR: 591253

Thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Communication and Information Sciences,
Master Track Data Science: Business And Governance,
at the School of humanities
of Tilburg University

Thesis committee:

Chris Emmery
Drew Hendrickson

Tilburg University
School of Humanities
Department of Communication and Information Sciences
Tilburg center for Cognition and Communication (TiCC)
Tilburg, The Netherlands
July 2017

Preface

In front of you lies the master thesis “Predicting job transition using publicly available professional profiles”, which is based on a method to predict job transition using publicly available data. It has been written to fulfill the graduation requirements of the Tilburg School of Humanities(TSH) at Tilburg University. I was engaged in researching and writing this thesis from January to July 2017. The completion of this thesis would not have been possible without the support and encouragement of the following people. I would like to thank 8Vance, for supplying the data, a workspace, and a constant stream of motivational "vlaai". I would like to especially thank Sabrina for her supervision and guidance throughout the project. Furthermore, I would like to thank Chris for his supervision and feedback throughout the thesis. Finally, I would like to express my love and appreciation for all support by my family and friends. The coffee breaks with my classmates from the DSBG master helped me through this, their hard work and dedication provided me with much motivation. Lastly, I would like to thank M.P., who provided me with support and inspiration when I needed them.

Evie Izeboud.

Tilburg, July 2017

Abstract

Prediction of job duration is an important problem in recruiting and human resources. Limitations of studies to date are the sole analyses of data from an internal company perspective and the strong focus on employee retention. This study adds to the existing literature by presenting models that can be used to predict job duration from an external perspective, investigating the ability to predict job duration using publicly available data. Furthermore, several different implementations of regression and classification were compared to determine which models suits this type of data best. These were trained on a subset of data originating from a selection of 65 million publicly available person profiles collected through web scraping. Features considered in this study are mean job duration (excluding the last completed job), total years worked and industry. Industry was excluded as a feature based on primary feature selection were industry did not show any predictive capability, this finding is in contrast with previous literature. The classification models all performed above baseline, all models had a very similar performance with predictive accuracy around 67%. The regression models also performed better than baseline, they all predicted some variance of the job duration, around 25%. Therefore, results for both tasks indicate that publicly available data has predictive merit for job duration. Selecting the best models was difficult, since the models yielded similar performance. Unfortunately all predictions were below practical relevance thresholds. Capturing a broader set of characteristics could significantly improve prediction accuracy, and help reach practically relevant levels. This could be achieved by combining publicly available external data with other data sources.

keywords: Human resource analytics, prediction, job duration, publicly available data, machine learning

Contents

1. Introduction

This section provides a short background for the study in subsection 1.1, the research questions in subsection 1.2, and the structure of this thesis in subsection 1.3.

1.1 Background

> *"The first step towards getting somewhere is to decide that you are not going to stay where you are."*
>
> – J.P. Morgan

The competitive and dynamic nature of the job market as well as personal goals and preferences lead individuals to change jobs at some point in their lives. Moving to a new job, however, is not an easy decision, and may depend on many factors. Boockmann and Steffes (2010) showed that currently more than fifty percent of all new employment relationships end within two years. This high employee turnover poses a problem for new and current employers. A number of trends (e.g., globalization, increase in knowledge work, accelerating rate of technological advancement) are believed to be responsible for this high turnover (Holtom, Mitchell, Lee, & Eberly, 2008), and make it vital that firms acquire and retain human capital.

Companies and recruiters often face the problem of deciding which individuals to approach for hiring. Whilst there are many ways for recruiters or companies to find and select individuals that have the right requirements using on-line and off-line curricula vitae, not all suitable candidates are necessarily willing to leave their current employment. It would be insightful for potential employers to know what individuals are likely to accept a new job, and how long that individual is likely to stay with them. If they can predict which individuals are likely to switch jobs, current and potential employers can consider whether to spend time and money on retaining, attracting or training those individuals. For individuals, it could be insightful to know when people with similar profiles transitioned to a new job, it might help them to decide when to take the next step in their career. Similarly, these models might be adapted to predict what type of job a person will transfer to, and even what skills they need for a certain type of job. Thus, due to all these different interests, the issue of employee turnover has been addressed extensively in the HR literature (see Section 2. Related work).

Prediction of job retention could be considered part of the field of Human Resource (HR) Analytics. In the HRA literature there is often a distinction between voluntary turnover and

involuntary turnover (Holtom et al., 2008). Both types of turnover have different types of causes, but both are relevant to this study. Therefore, when referring to turnover in this study, both types of turnover are taken into account. HR Analytics is a relatively new domain in the HR field, which aims to enable organizations to use descriptive, visual, and statistical analyses of data related to HR processes to establish business impact and facilitate data-driven decision-making, as explained by Marler and Boudreau (2017).

In their literature review Marler and Boudreau (2017) found little scientific evidence that newly available online public data sources are being used to guide decision-making in the field of HR analytics, despite evidence of a growing interest in this innovation. Marler and Boudreau (2017) selected 14 articles based on meeting scientific quality criteria from an initial population of 60 articles, ultimately only 4 involved empirical analyses of HR Analytics. The authors concluded that there is a growing need for more scientific research in this field. HR predictive analytics is an evolving application field of HR analytics. As stated by Mishra, Lama, and Pal (2016) an important part of HRPA is "Predictive Retention Modelling: Identify high-risk employees, build profiles, predict vacancies and leadership needs, and understand how risk is distributed throughout the organization"(p. 34). As can be seen from this definition, these predictions are often conducted from an internal company perspective. Missing from the literature however, are predictive analytics from an external perspective. This study will attempt to fill this research gap. To summarize: the goal of this study is to use machine learning to create a predictive model for job duration of individuals using only static publicly available information scraped from the web.

A practical application of a predictive model like the one in this study might be useful for a variety of practical implementations. An example of this is that it could complement digital recruitment platforms. For recruiters and hiring companies it is very relevant to know which candidates are likely to switch jobs. For individuals it might be insightful to know when people with similar career profiles change jobs, and when they should consider doing so themselves. It is unlikely that the models from this study can be directly applied in a practical setting, but this study might be used as a stepping stone for future implementations.

1.2 Research questions

The first research question of the thesis is formulated as follows:

RQ1: *Is it possible to predict job transition in individuals using only publicly available data?*

If the answer to research question 1 is anything other than no, a second research question can be answered:

RQ2: *What is the best model to predict which individuals are likely to switch jobs using only publicly available data?*

1.3 Structure

The outline of the remainder of the thesis is as follows. Section 2 (Related work) contains information on relevant work from previous studies, reviewing what has and has not been reported in the literature, justifying the added value of the current study. The section starts with an overview of the field of human resource analytics. Secondly, machine learning algorithms used for human resource analytics and specifically job prediction related applications will be discussed. Section 3 (Method) will describe the dataset and the experimental procedure in sufficient detail for other researchers to replicate the study. A description of the dataset will be provided. Furthermore, the pre-processing of the data will be discussed and information about feature selection will be provided. The experimental procedure will be explained, what tasks were performed, which algorithms were used, which parameters were chosen and why. The parameter optimization will be explained and implementation methods will be discussed. Lastly, the evaluation criteria, the evaluation scheme and error measures will be provided in the Method section. Section 4 (Results) will discuss the empirical results. The results from analysis of the differences between industries are discussed. Results from the classification task will be discussed as well as the results from the regression task. In section 5 (Discussion) the research question will be answered, and directions for future research will be given. The results will be discussed and interpreted. Suggestions for future research will be given and to conclude, the research questions will be repeated and answers to them provided by combining the results obtained with a very brief summary of how they can be placed in the context of existing research.

2. Related work

This section provides a context of related work and explains how this work relates to the literature. The section starts with an overview of the field of human resource analytics in

subsection 2.1. In subsection 2.2, machine learning algorithms used for human resource analytics and specifically job prediction related applications will be discussed.

2.1 Human resource analytics

Employee turnover has drawn researchers' and human resource experts' attention due to the associated cost of employee turnover that impacts the operational capabilities and organizational budget (Zhu et al., 2016). Nevertheless, most organizations seem to lack a consistent view of the workforce and thus need HR analytics to perform workforce optimization and to produce better "Return On Investment" (ROI) (Bassi, 2011).

Many HR analytics studies are based on internal company data and employee surveys (Collini, Guidroz, & Perez, 2015; Heponiemi, Kouvonen, Virtanen, Vänskä, & Elovainio, 2014; Y.-H. Huang et al., 2016; Jordan & Troth, 2011; Yousef, 2017). For example, the study by Collini et al. (2015) used employee surveys and linked those to turnover rates gathered from internal company records. However, this internal data is not always available for a third party, for example in the case of recruitment of new employees. Companies looking to attract new employees often have to rely on publicly available external data. The recruiting systems commonly used by companies, make use of social networks such as LinkedIn, Facebook, Twitter, Xing (Zide, Elman, & Shahani-Denning, 2014). Davison, Maraist, and Bing (2011) report that information on business-oriented social networking platforms such as LinkedIn or Xing are more accurate than social media platforms like Facebook, as people in the same network can view and verify the information provided.

Griffeth, Hom, and Gaertner (2000) conducted a meta-analysis of antecedents and correlates of employee turnover. The main predictors for employee turnover were: job satisfaction, organizational commitment,job search, comparison of alternatives, withdrawal cognitions, and quit intentions. Like the studies mentioned above,these are all internally measured through questionnaires. However, they also found evidence for an external factor that can predict employee turnover: alternative job opportunities. As CBS (2017) showed, alternative job opportunities are highly dependent on the industry a person is working in. Furthermore, in the study by Paparrizos, Cambazoglu, and Gionis (2011) evidence was found for industry as an important predictor for predicting future employers. Therefore, differences between industries will be investigated as a feature for predicting employee turnover.

2.2 Machine Learning algorithms

With increased digitalization, attention has started growing for e-recruitment, a recruitment process based on information publicly available on the web (Boudreau & Cascio, 2017; Mishra & Lama, 2016; Thompson, Braddy, & Wuensch, 2008). Mishra and Lama (2016) explained that when human resource data is assessed, different methods can be used to extract knowledge. Predictive analytics are known to increase accuracy of automated decision making, and combining data mining with advanced predictive techniques has provided more understanding in HR. For example, Mishra et al. (2016) applied customer churn models on employee turnover data to create predictive employee turnover models. Their paper demonstrates that machine learning techniques can be used to build reliable and accurate predictive models for employee turnover.

In this subsection, multiple classification and regression algorithms will be discussed. The justification for trying different algorithms instead of just one can be found in the "no free lunch" theorem for optimization stated by Wolpert and Macready (1997). This theorem has a highly theoretical and mathematical basis, but in practice it can roughly be interpreted as "no search algorithm by definition outperforms any other algorithm". From that it follows that one should consider multiple different algorithms and select the one with the best performance. It is certainly possible that some algorithms perform equally well and there is a tie. In that case one must take into account Occam's razor, a principle stated by the philosopher William of Ockham. This principle chooses simplicity over complexity: of two competing theories or models, the simpler explanation is to be preferred. However, it is difficult to objectively measure simplicity therefore mean training time in the cross-validation procedure will be used as a proxy for simplicity. In the text below the algorithms that describe job duration and/or employee turnover in the scientific literature will be discussed. All algorithms mentioned in this literature were investigated and will be explained briefly, unless specified differently.

Zhu et al. (2016) performed a study that predicted employee turnover using time series forecasting techniques. The study included longitudinal demographic metrics such as payroll category, hired date, termination date, age, years of service, gender, and job classification. Various time series forecasting models for predicting employee turnover were tested and optimal models for turnover forecasts were identified. Interestingly, Zhu et al. (2016) created a summary of previous research on employee turnover forecast (see Table 10). From their meta-analysis

by Zhu et al. (2016), it can be deducted that employee turnover predictions typically rely on linear regression and logistic regression and to allow comparison, this study will include those methods.

2.2.1 Linear regression. In statistics, linear regression is an approach for modeling the relationship between a continuous dependent variable and one or more explanatory variables, or features in machine learning. If there is more than one explanatory variable, it is referred to as multiple linear regression. Stanton (2001) study on publications of Sir Francis Galton and Karl Pearson revealed that Galton's work on inherited characteristics of sweet peas led to the initial conceptualization of linear regression. Subsequent efforts by Galton and Pearson brought about the more general techniques of multiple regression and the product-moment correlation coefficient. In linear regression a line is fitted to the data as to minimize the sum of squared residuals. Linear regression analysis is a widely used statistical technique with all types of applications, including employee turnover (Bluedorn, 1982; Collini et al., 2015; Thaden, Jacobs-Priebe, & Evans, 2010).

2.2.2 Logistic regression. Logistic regression is a supervised regression model where the dependent variable (DV) is categorical. Logistic regression is also referred to as logit regression, or logit model (Freedman, 2009). In traditional logistic regression, the outcome variable is binary; there are only two classes. Logistic regression was developed by statistician David Cox (1958). Although it is an old method, logistic regression was used in many recent studies predicting employee turnover (Y.-H. Huang et al., 2016; Y.-h. Huang et al., 2014; Li, Lee, Mitchell, Hom, & Griffeth, 2016; Stanley, Vandenberghe, Vandenberg, & Bentein, 2013; Tews, Stafford, & Michel, 2014; Vardaman, Taylor, Allen, Gondo, & Amis, 2015).

Paparrizos et al. (2011) addressed the problem of predicting future employers for individuals. They formulated their recommendation as a supervised machine learning model. They used publicly available profiles from the web to obtain information about past job transitions and job-related features to predict an individual's next employer. They used 3 different samples, the first including people form the top 100 universities and top 100 companies, the second containing people from the top 100 companies, and the third containing only people from the top 25 companies. For each sample they predicted class is an institution among the most frequent 25 companies in the full data. For these 3 set ups they reached an accuracy of 67%, 78% and 86% respectively, with a baseline accuracy of 15%. The most important predictors in their study

were company title, industry and industry type(public or private). For this prediction they used a combination of the decision tree and Naive Bayes algorithm: the Decision tree/Naive Bayes hybrid classier (DTNB). Therefore, this study uses a decision tree and Naive Bayes as classifiers. Using a DTNB hybrid classifier was unfortunately not possible due to lack of an an available implementation of this algorithm and too limited time to create a implementation.

The Naive Bayes classifier is a supervised machine learning technique that uses the concept of probability to classify new entities. In the statistics and computer science literature, Naive Bayes models are known under a variety of names, including simple Bayes and independence Bayes (Hand & Yu, 2001). This method applies Bayes' theorem with the assumption of strong conditional independence assumptions between the features. In lay men's terms Naive Bayes considers every feature to be unrelated to the other features. Naive Bayes has a surprisingly competitive performance in classification (Lewis, 1998), considering that the conditional independence assumption on which it is based, is rarely true in real world applications. One of the advantages of the Bayesian classifier is that it is applicable in many different domains and situations.

2.2.4 Decision tree. The decision tree is a supervised machine learning technique that uses a predictive tree model to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). decision trees can be used for both regression and classification. Decision trees are computationally fast, make no statistical assumptions, and can handle data that are represented on different measurement scales (Pal & Mather, 2003). However, decision trees do have a tendency to over-fit on the training data (Hastie, Tibshirani, & Friedman, 2002).

2.2.5 Random Forest. To decrease the risk of over- fitting a decision tree, random forest is often used as an alternative (Hastie et al., 2002). Random forest, also known as random decision forest, was first developed by Ho (1995) and is a supervised method for classification, that operates by constructing a collection of decision trees in the training phase and outputs the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

2.2.6 Support Vector Machines. Saradhi and Palshikar (2011) used Linear Support Vector Machine (SVM) in their employee turnover prediction. Linear SVM is a supervised learning algorithm used for binary classification. An SVM model uses examples as points in

space, mapped so that the examples of the separate categories are divided by a hyperplane that is as wide as possible. Originally SVM was intended for binary classification purposes. However, a version of SVM for regression was proposed by Drucker, Burges, Kaufman, Smola, and Vapnik (1997).

Sexton, McMurtrey, Michalopoulos, and Smith (2005) used neural networks combined with a modified genetic algorithm to build a turnover prediction model. This method is called Neural Network Simultaneous Optimization Algorithm (NNSOA).The NNSOA was shown to perform well for optimizing a NN while simultaneously eliminating unnecessary weights in the NN structure during the training process for an employee turnover problem. The NNSOA was able to predict turnover with an average accuracy of 99.3%.

2.2.7 Neural Network. A Neural network neural networks can consist of one or more hidden layers of artificial neurons. Traditionally neural networks consisted of only several hidden layers. However, as a result of advances in hardware, techniques and data, the term deep learning has emerged (Schmidhuber, 2015). In deep learning, multiple processing layers are used to learn representations of data with multiple levels of abstraction. Advances in hardware being GPU-accelerated computing: the use of a graphics processing unit (GPU) together with a CPU to accelerate deep learning applications. Advances in techniques include better weight initialization from unsupervised techniques. And with large public datasets are becoming more common, even more personal data is becoming more widely available. Most machine learning methods require personal data to be solved more accurately. The data in this study only contains public data, therefore it will be interesting to see how this influences both the regular machine learning methods as the deep learning methods. For this study the deep learning library Keras with a Tensorflow back-end was used, see subsection 3.6 for more details.

3. Method

This section describes the methods and experimental setup for this study, in a way that enables replication. In subsection 3.1 the terms which are used throughout this thesis will be defined. A description of the dataset will be provided in subsection 3.2. The pre-processing of the data will be discussed in section 3.3. In subsection 3.4, the experimental procedure will be explained, what task was performed, which algorithms were used, which parameters were chosen and why. The implementation methods will be discussed in section 3.5. In section 3.6, Evaluation criteria,

the evaluation scheme and error measures will be provided.

3.1 Description of the dataset

For this study, a dataset was provided by the company 8Vance[1] in Venlo. The dataset consists of a large sample of publicly available person profiles that were extracted from the web (about 65 million), containing information about job transitions and associated meta-data. The dataset is structured and stored in an Apache Cassandra database created by Lakshman and Malik (2010). Apache Cassandra is an open source distributed database management system and according to Chebotko, Kashlev, and Lu (2015) it is a leading distributed database of choice when it comes to big data management.

Subsets of profiles where randomly sampled from the dataset to obtain a training and test subset. See Figure 1 for a visual representation of how the subsets of the data were created. The dataset contains Dutch, English and German publicly available person profiles. The profiles contain information about the employees' professional experiences. Nominal features that were measured on a job-level (variables relating to the employment): function type, company name, industry type, start-date and stop-date. Features that were measured on an individual level (variables relating to the individual and the total career of that individual): total years worked and mean job duration (see 3.3 Data pre-processing for more information on how these features were constructed).

3.2 Data pre-processing

The dataset was already cleaned and pre-processed by the company. There were 5 profiles in the dataset that reported a total work experience of more than a 60 years, which is considered highly unlikely, therefore these profiles were omitted from the data. From this dataset a subset of 10.000 profiles that met specified requirements were selected:

- profiles that contain between two and ten non-current jobs
- all jobs included information about the start-date, stop-date, total years worked and industry type
- less than 3 months between each job

[1] www.8vance.com/

- last job duration at least 6 months

Profiles with between two and ten non-current jobs were sampled as to makes sure freelancers were excluded, but at least two observations per person were present so a mean of previous job durations can be calculated. The second requirement was set to make sure there was no missing data. Although it is possible to make prediction in a dataset that suffers from missing data, complete profiles were preferred. The third requirement is based on the fact that it is of interested that people transition between jobs, and not just quit a job without starting a new one. However, it is not uncommon that people take a period of a few weeks as a small vacation period between jobs, or finding a new job might take a few weeks. Therefore, a boundary of less than 3 months was chosen. The last requirement is to make sure that the job duration of the last non-current job was not too short.

The feature total years worked was created by the company, it is based on the duration of all jobs, taking into account any overlap that may occur when individuals had more than one job at the same time. The start-date and stop-date were used to create a new feature of job duration in years. The job durations in its turn were used to calculate the mean job duration for all non-current jobs apart form the last non-current job for each individual. The last job was excluded from this mean so that the last job could be used as an outcome variable.

3.3 Features

As mentioned before Paparrizos et al. (2011) found evidence for industry as an important predictor for predicting future employers. Therefore, we investigated differences between industries as a feature for predicting employee turnover. Because there are more than 500 industries in the dataset, we selected a subset containing people that had jobs in the top 10 most reported industries. The industries, ranked from highest to lowest frequency: "IT and Services", "Hospital & Health Care", "Financial Services", "Marketing and Advertising", "Construction", "Education Management", "Computer Software", "Retail", "Higher Education", and "Banking".

The American Bureau of Labor Statistics reported on the duration of employment relationships with a single employer for all jobs started from age 18 to age 48 in 1978-2012 (BLS, 2015). According to the U.S. Bureau of Labor Statistics the average person born in the latter years of the baby boom (1957-1964) held 11.7 jobs from age 18 to age 48. They showed that the average duration of employment is shorter for people at the beginning of their career, than

for people at the end of their career. Therefore it is a logical step to take into account how far along people are within their career, represented by the variable total years worked.

In a study by Khatri, Fern, and Budhwar (2001) it was found that job-hopping attitudes of employees is one of the most important predictors for employee turnover. Despite the fact that we do not have information about job-hopping attitudes, we do have information about job-hopping behavior. We will use the mean job duration of all previous non-current jobs (excluding the last non-current job) as a proxy for job-hopping behavior. The last non-current job was excluded because it is the dependent variable in this study.

All three features were used for both the classification and the regression task. Descriptive statistics for all features except industry can be found in Table 2. Because industry is a nominal feature, in Table 3 frequencies for every industry in the full subset of a 100.000 profiles (train and test data) can be found.

3.4 Experimental procedure

Two tasks were performed, a regression and a classification task. The regression task was performed to predict a continuous outcome: how long (measured in years) a person stayed at their job. The classification task was performed to predict a categorical outcome: whether people stay for less or more than two years at their job. We classified individuals into two categories: a short stay (<2 years) and into a long-stay(>2 years) group based on their non-current job durations. To improve comprehensibility and because the distribution of mean job durations is not symmetrical, two years was used as a cut-off point. The cut-off point of two years is based on the median last job duration for the training data, which is 2.4 years. This resulted in two balanced groups: 52% long-stay versus 48% short-stay. This in accordance with the findings of Boockmann and Steffes (2010) who reported that more than 50% of employees leave within two years. The features used for both tasks are: industry type, total years worked and mean job duration (excluding the last job).

Based on the literature discussed in the related work, five algorithms were selected for each task. For the classification task, five classifiers were tested. The classifiers considered are Logistic Regression, Decision Tree, Random Forest, Naive Bayes, and a Neural network. For the regression task, five algorithms were tested. The regression algorithms considered are linear regression, Decision Tree, Random Forest, Linear SVM, and a Neural network. Hyperparameters

were chosen using hyperparameter optimization, which will be discussed in the next subsection.

3.5 Hyperparameter optimization

In the context of machine learning, hyperparameter optimization is "the problem of optimizing a loss function over a graph-structured configuration space" (J. S. Bergstra, Bardenet, Bengio, & Kégl, 2011). Practically, hyperparameter optimization can be defined as choosing a set of hyperparameters for a learning algorithm, usually with the goal of optimizing a measure of the algorithm's performance on an independent data set. The best values of hyperparameters are chosen by minimizing a certain criteria, for example, error classification on a validation set.

A widely used strategy for hyperparameter optimization is a combination of grid search and manual search (Bardenet, Brendel, Kégl, & Sebag, 2013; Larochelle, Erhan, Courville, Bergstra, & Bengio, 2007; LeCun, Bottou, Bengio, & Haffner, 1998). However, grid searches suffer from the curse of dimensionality because the number of possible outcomes grows exponentially with the number of hyper-parameters (J. Bergstra & Bengio, 2012). The curse of dimensionality entails that if one has a high-dimensional feature space with each feature having a number of possible values, an enormous amount of training data is required to ensure that there are several samples with each combination of values. Manual search is used in this process to decrease the number of possible hyperparameters by selecting a specified subset of the hyperparameter space of a learning algorithm. The combination with manual search helps speed up the process, but manual tuning requires considerable expertise which leads to poor reproducibility.

J. Bergstra and Bengio (2012) proposed random search as an alternative keeping the advantages of implementation simplicity and reproducibility of pure grid search. J. Bergstra and Bengio (2012) showed that "random search has all the practical advantages of grid search (conceptual simplicity, ease of implementation, trivial parallelism) and trades a small reduction in efficiency in low-dimensional spaces for a large improvement in efficiency in high-dimensional search spaces". Consequently, random search was used for the hyperparameter optimization in this study.

3.6 Implementation methods

We will evaluate and compare multiple methods based on theory and performance. Analyses will be conducted mainly in Python, supplementary analyses will be done in R. To implement the machine learning algorithms (except the Neural Network), Scikit-learn was used (Pedregosa et

al., 2011). Scikit-learn can be used to implement many well known machine learning algorithms, while maintaining an easy to-use interface tightly integrated with the Python language.

For the Neural network models a different library a Python deep learning library was used: Keras (Chollet et al., 2015). Keras allows for easy and fast prototyping, supports both convolutional networks and recurrent networks, and runs seamlessly on CPU and GPU. Within Keras a sequential model will be used, which is a linear stack of layers. Keras works on top of three types of frameworks: TensorFlow, CNTK or Theano. For this study a Tensorflow backend was used. Tensorflow is an interface for expressing machine learning algorithms, and an implementation for executing such algorithms (Abadi et al., 2015).

3.7 Evaluation criteria

3.7.1 Evaluation scheme. Training an algorithm and evaluating its statistical performance on the same data often leads to over-fitting and as a result an overestimation of performance on new data. There are multiple methods that can be used to prevent over-fitting and get a more realistic performance estimate, including, but not limited to k-fold cross-validation. Validation is the process of splitting the data into a training and a validation part, were the training data is used for training and the validation data for evaluation.

Cross-validation is a technique that can be used to assess how well the predictive ability of an algorithm can be generalized to an independent dataset (Arlot, Celisse, et al., 2010). The dataset is randomly split into k mutually exclusive subsets (the folds) of approximately equal size. Each subset is used as the "test" set once, and used to evaluate the model that was fit using all other subsets as training data. this process is repeated so that all folds are used as the "test" set once. The cross-validation estimate of accuracy is the overall number of correct classifications,divided by the number of instances in the dataset.

In a study by Molinaro, Simon, and Pfeiffer (2005), it was shown that for real world datasets k-fold cross-validation is superior to leave one-out cross-validation and bootstrapping and has a lower bias with respect to accuracy estimation and model selection. Kohavi et al. (1995) demonstrated that ten-fold cross-validation is often better, even if the data allows for more folds. Therefore we will use ten-fold cross-validation as an evaluation scheme. The evaluation scheme that is used to estimate the accuracy of the models used for this research will therefore be 10-fold cross-validation.

3.7.2 Error measures. To evaluate our method, we will split the data into a training set and test set, using the training set for cross-validation. Based on the cross-validation results, the best model/models will be selected and the test set will evaluated on performance. For this procedure an error measure is needed to quantify the performance of the models. The error measure and baseline method (to which we will compare our method) differ for the different tasks.

The classification models will be evaluated by predictive accuracy: this refers to the ability of the model to correctly predict the class label previously unseen data. The accuracy will be measured as the percentage of examples correctly classified by the classifier. For the classification task the outcome variable is binary, "1" for people that stayed for less than two years at their last job, "0" for people that stayed for two or more years. The majority baseline method will be used as the baseline model for the classification task. Majority baseline is a model were the majority class is predicted as an outcome for all individuals.

Because there are no studies in the current literature that try to predict job duration using publicly available professional profiles, it is not possible to use a result from the literature as a threshold. Consequently, an arbitrary threshold was specified that would be practically relevant to the dataset owner, a threshold of 90% accuracy was chosen. For companies like the dataset owner, it is necessary to have such a high accuracy to be effective and appealing to customers when applied in practice. To illustrate this: we know that if one randomly guesses whether someone will stay at their current job for more than two years, they have a 52% probability of guessing it correctly if we just assume everyone leaves within two years. Let us assume that recruiters have experience and recruitment strategies that help them to identify what individuals are likely to leave soon, that are better than random guessing (although some might beg the differ). A model that could benefit recruiters should therefore do more than just perform slightly above baseline, to convince a recruiter to use a model like the one proposed in this study.

The regression models will evaluated on goodness of fit using the R^2 value. The best fitting model is selected based on a highest R^2 value in the cross-validation procedure. The goodness of fit R^2 value, also known as the coefficient of determination, can be defined as the percentage of variance in the outcome variable that is explained by the model. As a baseline model mean prediction will be used, for every individual the mean job duration for the training set will be predicted. Mean prediction automatically leads to a R^2 of zero. Because the first research

question is "Is it possible to predict job transition in individuals using only publicly available data?", any value for R^2 that is not zero indicates that the model explains some variance from the job duration. Therefore, if R^2 is not zero, we can conclude that it is (partially) possible to predict job transition with publicly available data. The higherR^2 the better the predictive abilities of the model, with one indicating a perfect predictive accuracy.

Because there are no studies in the current literature that try to predict job duration using publicly available professional profiles, it is not possible to use a result from the literature as a threshold. Consequently, a threshold was specified that would be practically relevant to the dataset owner. For a regression model to be practically relevant, a threshold of .80 for the R^2 was chosen.

Because there are no studies in the current literature that try to predict job duration using publicly available professional profiles, it is not possible to use a result from the literature as a threshold. Consequently, an arbitrary threshold was specified that would be practically relevant to the dataset owner, a threshold of .80 for the R^2 was chosen. For companies like the dataset owner, it is necessary to have such a high R^2 to be effective and appealing to customers when applied in practice. However, it is lower than the threshold for the classification task, since means in the case of regression it is hard to exactly make a perfect prediction for every instance and small errors are generally acceptable. To illustrate this: Let us assume we have an average person with a job duration which is exactly equal to the mean job duration: 2.41 years. If a prediction is made with 80% accuracy, it means there predicted job duration is expected to be somewhere between 1.93 years and 2.89 years. This a difference of half a year more or less, which could be relevant for effective recruitment.

4. Results

In subsection 4.1 results from the differences between industries will be investigated discussed. Results from the classification task will be discussed in section 4.2, and the results from the regression task will be discussed in section 4.3.

4.1 Feature selection

It is only relevant to investigate industry as a feature in the model, if there are differences in job durations between different industries. An Analysis of Variance (ANOVA) was performed to investigate whether people in different industries on average have a different job duration.

The ANOVA showed that the effect of industry on the duration of employment was significant, $F_{(9,25964)} = 22.72, p < .001$, indicating differences in job durations between industries. Post-hoc comparisons using the Tukey Honest significant Differences (HSD) test showed 17 significant differences in means between industries ($p < .05$), out of a total of 45 post-hoc comparisons. For example, there was a difference in mean job duration between the Banking industry and the IT and Services industry, with ($p < .001$). Because so many post hoc comparisons were performed, a visual representation of the differences in mean job durations between industries can be found in 2. For a complete overview of all post hoc results, see the Appendix.

Secondly, for all three potential predictors (industry, years worked and mean previous job duration) separate models were made based on the training data and using cross-validation, to asses which features can be used to predict job duration. Results of these analyses can be found in Table 4 for the classification task and table 5 for the regression task. In contrast with what was reported in the literature, industry does not show any predictive ability and scores around baseline. Years worked and mean previous job duration both do show predictive ability. Therefore, these last two features were selected for the classification and the regression models.

4.2 Classification results

To investigate if it is possible to predict job duration using external features, five classifiers were tested using the two selected features. As a baseline model the majority baseline of the outcome variable in the training data was used. In the majority baseline model, the majority class, "short-stay" is assigned as an outcome for all individuals. Therefore baseline model has an accuracy of .52, as reported before 52% of the people stayed at their job for less than 2 years.

Results of the classification task can be found in Table 6. Included in the table are the hyperparameters that were selected during the hyperparameter optimization. As can be seen, all classification algorithms perform better than baseline in cross-validation. This indicates that it is possible to (partially) predict job duration using only external features. Choosing the best model is difficult, because the performances of all algorithms do not statistically differ. Because of equal performance in the cross-validation, training time was considered. However, training times did not differ much between the models, making it impossible to select a truly "best" model, so no specific model was selected. To check how well the models generalize to new data, models were also applied to the test data. This yielded a result equal to that of

the cross-validation performance, which indicates that the models generalizes well to new data. However, the performance of these models is not very practically relevant, since the threshold was set at 90% accuracy.

Since the models had performance scores very close to one another, a correlation matrix of the test predictions was made to see how often the models with similar performances make the same predictions. These correlations are based on the prediction made on the test data. As can be seen in Table 7 The Decision Tree and Random Forest model correlate for 99%, this means that 99% of the predictions are the same. All other models show overlap significantly higher than expected by chance. This explains why they have similar performances, they often make the same predictions.

4.3 Regression results

To investigate if it is possible to predict job duration using external features, seven algorithms were tested using a feature vector including continuous and binary dummy features. The baseline model uses the mean of the outcome variable: the duration of each job, which has a mean of 2.41 years. The baseline model has, by definition, an R^2 of 0.

Results of the regression task can be found in Table 8. As can be seen, all regression algorithms perform better than baseline in the cross-validation. This indicates that it is possible to (partially) predict job duration using only external features. All the models had a similar performance, subsequently all models were applied to the test data for comparison purposes. Selecting a model based on training time was difficult, since training times did not differ greatly. As can be seen in the table, all models only showed minor differences between the cross-validation and test performance indicating that all models generalize well on new data. However, the performance of all models is not practically relevant, since the threshold was set at 75% explained variance.

In accordance with the classification task, the regression models also had performance scores very close to one another, a correlation matrix of the predictions for the test data was made to see how often the models with similar performances make the same predictions. As can be seen in Table 9, surprisingly the Linear regression and the Neural Network correlate for 98%, this means that 98% of the predictions are the same. All other models show overlap significantly higher than expected by chance. This explains why they have similar performances, they often

make the same predictions.

5. Discussion and conclusions

In this section a general discussion of the research will be accompanied by recommendations for further research. In Subsection 5.1 the results of section 4 will be discussed and interpreted. In Subsection 5.2 suggestions for future research will be given. Lastly, a short conclusion will be presented in subsection 5.3.

5.1 Interpretation of results

The main goal of this study was to create a predictive model for job transition of individuals using only publicly available information. Machine learning algorithms were used to model and evaluate the expected duration of employment.

All classifiers performed similar and better than the baseline, the training times did not differ greatly. Therefore it was not possible to select the best model for this classification task. Similarly, for the regression task performance was roughly equal and better than the baseline, and the training times did not differ greatly. All models were therefore also applied to the test data for comparison, these models all had a test performance close to the cross-validation performance. Unfortunately performance was still not good enough to be practically relevant for both classification and regression.

It was surprising that all models for both tasks had very similar performances. This is possible related to the fact that none of them had a very good performance, performing just above baseline. The models showed high correlations with one another, with a correlation between .84 and .99. The performance of the two features, years worked and mean job duration did not perform much better than the model with only mean job duration. This could indicate that both variables explain the same variance in the outcome variable.

Although the results show that it is possible to predict job duration classification better than baseline, only 2/3 of the predictions were correct, leaving 1/3 of predictions incorrect. The regression model yielded similar results, only a small improvement from baseline that is not practically relevant. This might be explained by the fact that the dataset only consisted of publicly available external data. When an individual decides whether to stay or leave a job, many factors play a role. In a later review of the literature on employee turnover Ongori (2007) reported four different types of factors: organization-wide factors, immediate work environment

factors, job-related factors, personal factors. In the current study only job-related factors were taken into account. In a meta-analysis on employee-turnover byGriffeth et al. (2000), it was discovered that internal information were the most dominant predictors of employee-turnover, this includes features such as job satisfaction, organizational commitment,job search, comparison of alternatives, withdrawal cognitions, and quit intentions with regards to the current job. This study did not take into account any of these features, because this information was not publicly available. A possible solution for this might be to combine publicly available data with internal data.

The fact that all models performed similary and had similar training times, made it impossible to objectively select the best classifier for either task. A limitation of the classification is that a continuous variable, duration in years, was dichotomized to a binary variable representing shortstay (<2 years) and longstay (>2 years). Fedorov, Mannino, and Zhang (2009) demonstrated that a consequence of dichotomizing a continuous variable is a loss of information. However, most literature on employee turnover makes use of classification and not regression methods. Therefore it was decided to follow the existing research methodology and use a classification model and compare this with a regression model. Generally the models show very similar results.

5.2 Suggestions for future research

With the knowledge obtained through this study, other researchers could be inspired to use publicly available data as well. An important suggestion for future research is to try and combine different data sources to gain more insight into employee turnover. Ideally, one would combine data relating to all the four factors reported by Ongori (2007): organization-wide factors, immediate work environment factors, job-related factors, personal factors.

For the sake of workability this study has focused on the top 10 most reported industries. these ten industries showed only small differences in distributions of job durations. It might be interesting for future research to include more industries and to investigate differences between these industries.

Another suggestion for future research is to investigate the relationships between self-reported functions types and skills. Using the professional skills that are often provided in publicly available professional profiles it might be possible to predict what type of function an

individual is eligible for. Or it might be possible to specify what skills one needs to develop to obtain a certain function.

A practical application a predictive model like the one in this study might be useful for a variety of recruitment platforms. An example of this is that it could complement the Automatic Intelligent Matching Agent (AIMA) from 8vance, the company that provided the dataset. AIMA is a virtual career and recruitment assistant for job mobility and talent acquisition. For recruiters and hiring companies it is very relevant to know which candidates are likely to switch jobs. For individuals it might be insightful to know when people with similar career profiles change jobs, and when they should consider doing so themselves. To do this the platforms might need to improve the current model, for example by adding self-reported data related the factors mentioned by Ongori (2007): organization-wide factors, immediate work environment factors, job-related factors, personal factors.

5.3 Conclusions

In this section the research questions and the answers that were found will be discussed. The first research question is:

RQ1: *Is it possible to predict job transition in individuals using only publicly available data?*

The answer to this question is: Yes, it is possible to predict which individuals are likely to switch jobs using only publicly available data. Both the classification and regression models performed better than baseline in the cross-validation. The classification models improved the prediction to 67% accuracy, with a baseline of 52% accuracy. The regression models improved the baseline model, they explain around 25% of the variance in the test data, as opposed to a baseline of 0%.

Because the answer to research question 1 was not no, a second research question could be answered:

RQ2: *What is the best model to predict which individuals are likely to switch jobs using only publicly available data?*

The classification models all performed above baseline, all models had a very similar performance around 67%. The regression models also performed better than baseline, they all predicted some variance of the job duration, around 25%. Training times did not differ significantly between

different models for both task, consequently it was impossible to select the best model and answer research question one.

It can be concluded that this is an interesting lead for further research. If this model can developed further and improved, it might be very relevant for practical as well as scientific applications in the field of HR.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... Zheng, X. (2015). *TensorFlow: Large-scale machine learning on heterogeneous systems.* Retrieved from `http://tensorflow.org/` (Software available from tensorflow.org)

Arlot, S., Celisse, A., et al. (2010). A survey of cross-validation procedures for model selection. *Statistics surveys*, *4*, 40–79.

Bardenet, R., Brendel, M., Kégl, B., & Sebag, M. (2013). Collaborative hyperparameter tuning. In *International conference on machine learning* (pp. 199–207).

Bassi, L. (2011). Raging debates in hr analytics. *People and Strategy*, *34*(2), 14.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*(Feb), 281–305.

Bergstra, J. S., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. In *Advances in neural information processing systems* (pp. 2546–2554).

BLS. (2015). *Number of jobs held, labor market activity, and earnings growth among the youngest baby boomers: Results from a longitudinal survey.* Retrieved from `https://www.bls.gov/news.release/pdf/nlsoy.pdf`

Bluedorn, A. C. (1982). A unified model of turnover from organizations. *Human relations*, *35*(2), 135–153.

Boockmann, B., & Steffes, S. (2010). Workers, firms, or institutions: What determines job duration for male employees in germany? *ILR Review*, *64*(1), 109–127.

Boudreau, J., & Cascio, W. (2017). Human capital analytics: why are we not there? *Journal of Organizational Effectiveness: People and Performance*, *4*(2).

CBS. (2017). *Banen en vacatures naar bedrijfstak.* Retrieved 2010-09-30, from `https://www.cbs.nl/nl-nl/achtergrond/2017/07/banen-en-vacatures-naar-bedrijfstak`

Chebotko, A., Kashlev, A., & Lu, S. (2015). A big data modeling methodology for apache cassandra. In *Big data (bigdata congress), 2015 ieee international congress on* (pp. 238–245).

Chollet, F., et al. (2015). *Keras.* `https://github.com/fchollet/keras`. GitHub.

Collini, S. A., Guidroz, A. M., & Perez, L. M. (2015). Turnover in health care: the mediating effects of employee engagement. *Journal of nursing management*, *23*(2), 169–178.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, 215–242.

Davison, H. K., Maraist, C., & Bing, M. N. (2011). Friend or foe? the promise and pitfalls of using social networking sites for hr decisions. *Journal of Business and Psychology*, *26*(2), 153–159.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A. J., & Vapnik, V. (1997). Support vector regression machines. In *Advances in neural information processing systems* (pp. 155–161).

Fedorov, V., Mannino, F., & Zhang, R. (2009). Consequences of dichotomization. *Pharmaceutical Statistics*, *8*(1), 50–61.

Freedman, D. A. (2009). *Statistical models: theory and practice*. cambridge university press.

Griffeth, R. W., Hom, P. W., & 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*, *26*(3), 463–488.

Hand, D. J., & Yu, K. (2001). Idiot's bayes—not so stupid after all? *International statistical review*, *69*(3), 385–398.

Hastie, T., Tibshirani, R., & Friedman, J. (2002). The elements of statistical learning: Data mining, inference, and prediction. *Biometrics*.

Heponiemi, T., Kouvonen, A., Virtanen, M., Vänskä, J., & Elovainio, M. (2014). The prospective effects of workplace violence on physicians' job satisfaction and turnover intentions: the buffering effect of job control. *BMC health services research*, *14*(1), 19.

Ho, T. K. (1995). Random decision forests. In *Document analysis and recognition, 1995., proceedings of the third international conference on* (Vol. 1, pp. 278–282).

Holtom, B. C., Mitchell, T. R., Lee, T. W., & Eberly, M. B. (2008). 5 turnover and retention research: A glance at the past, a closer review of the present, and a venture into the future. *Academy of Management annals*, *2*(1), 231–274.

Huang, Y.-H., Lee, J., McFadden, A. C., Murphy, L. A., Robertson, M. M., Cheung, J. H., & Zohar, D. (2016). Beyond safety outcomes: An investigation of the impact of safety climate on job satisfaction, employee engagement and turnover using social exchange theory as the theoretical framework. *Applied ergonomics*, *55*, 248–257.

Huang, Y.-h., Lee, J., McFadden, A. C., Murphy, L. A., Robertson, M. M., & Zohar, D. (2014). The impact of safety climate beyond safety outcomes: Job satisfaction, employee engage-

ment and objective turnover rate. In *Proceedings of the 11th international symposium on human factors in organisational design and management, copenhagen, denmark.*

Jordan, P. J., & Troth, A. (2011). Emotional intelligence and leader member exchange: The relationship with employee turnover intentions and job satisfaction. *Leadership & Organization Development Journal*, *32*(3), 260–280.

Khatri, N., Fern, C. T., & Budhwar, P. (2001). Explaining employee turnover in an asian context. *Human Resource Management Journal*, *11*(1), 54–74.

Kohavi, R., et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (Vol. 14, pp. 1137–1145).

Lakshman, A., & Malik, P. (2010). Cassandra: a decentralized structured storage system. *ACM SIGOPS Operating Systems Review*, *44*(2), 35–40.

Larochelle, H., Erhan, D., Courville, A., Bergstra, J., & Bengio, Y. (2007). An empirical evaluation of deep architectures on problems with many factors of variation. In *Proceedings of the 24th international conference on machine learning* (pp. 473–480).

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Lewis, D. D. (1998). Naive (bayes) at forty: The independence assumption in information retrieval. In *European conference on machine learning* (pp. 4–15).

Li, J. J., Lee, T. W., Mitchell, T. R., Hom, P. W., & Griffeth, R. W. (2016). The effects of proximal withdrawal states on job attitudes, job searching, intent to leave, and employee turnover. *Journal of Applied Psychology*, *101*(10), 1436.

Marler, J. H., & Boudreau, J. W. (2017). An evidence-based review of hr analytics. *The International Journal of Human Resource Management*, *28*(1), 3–26.

Mishra, S. N., & Lama, D. R. (2016). A decision making model for human resource management in organizations using data mining and predictive analytics. *International Journal of Computer Science and Information Security*, *14*(5), 217.

Mishra, S. N., Lama, D. R., & Pal, Y. (2016). Human resource predictive analytics (hrpa) for hr management in organizations. *International Journal of Scientific & Technology Research*, *5*(5).

Molinaro, A. M., Simon, R., & Pfeiffer, R. M. (2005). Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, *21*(15), 3301–3307.

Ongori, H. (2007). A review of the literature on employee turnover.

Pal, M., & Mather, P. M. (2003). An assessment of the effectiveness of decision tree methods for land cover classification. *Remote sensing of environment*, *86*(4), 554–565.

Paparrizos, I., Cambazoglu, B. B., & Gionis, A. (2011). Machine learned job recommendation. In *Proceedings of the fifth acm conference on recommender systems* (pp. 325–328).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

Saradhi, V. V., & Palshikar, G. K. (2011). Employee churn prediction. *Expert Systems with Applications*, *38*(3), 1999–2006.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, *61*, 85–117.

Sexton, R. S., McMurtrey, S., Michalopoulos, J. O., & Smith, A. M. (2005). Employee turnover: a neural network solution. *Computers & Operations Research*, *32*(10), 2635–2651.

Stanley, L., Vandenberghe, C., Vandenberg, R., & Bentein, K. (2013). Commitment profiles and employee turnover. *Journal of Vocational Behavior*, *82*(3), 176–187.

Stanton, J. M. (2001). Galton, pearson, and the peas: A brief history of linear regression for statistics instructors. *Journal of Statistics Education*, *9*(3), 1–16.

Tews, M. J., Stafford, K., & Michel, J. W. (2014). Life happens and people matter: Critical events, constituent attachment, and turnover among part-time hospitality employees. *International Journal of Hospitality Management*, *38*, 99–105.

Thaden, E., Jacobs-Priebe, L., & Evans, S. (2010). Understanding attrition and predicting employment durations of former staff in a public social service organization. *Journal of Social Work*, *10*(4), 407–435.

Thompson, L. F., Braddy, P. W., & Wuensch, K. L. (2008). E-recruitment and the benefits of organizational web appeal. *Computers in Human Behavior*, *24*(5), 2384–2398.

Vardaman, J. M., Taylor, S. G., Allen, D. G., Gondo, M. B., & Amis, J. M. (2015). Translating intentions to behavior: The interaction of network structure and behavioral intentions in understanding employee turnover. *Organization Science*, *26*(4), 1177–1191.

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, *1*(1), 67–82.

Yousef, D. A. (2017). Organizational commitment, job satisfaction and attitudes toward organizational change: A study in the local government. *International Journal of Public Administration*, *40*(1), 77–88.

Zhu, X., Seaver, W., Sawhney, R., Ji, S., Holt, B., Sanil, G. B., & Upreti, G. (2016). Employee turnover forecasting for human resource management based on time series analysis. *Journal of Applied Statistics*, 1–20.

Zide, J., Elman, B., & Shahani-Denning, C. (2014). Linkedin and recruitment: How profiles differ across occupations. *Employee Relations*, *36*(5), 583–604.

Appendix

Table 1

Post-hoc Tukey test results: mean difference, Left and right bound of the 95% confidence interval, p value

	Difference	**Left bound**	**Right bound**	**p**
Hospital & Health Care-IT and Services	0.164	-0.005	0.334	0.067
Financial Services-IT and Services	0.108	-0.021	0.237	0.192
Marketing and Advertising-IT and Services	-0.519	-0.719	-0.319	<.001
Construction-IT and Services	0.043	-0.185	0.272	1
Education Management-IT and Services	0.052	-0.273	0.376	1
Computer Software-IT and Services	-0.022	-0.196	0.152	1
Retail-IT and Services	-0.204	-0.344	-0.063	<.001
Higher Education-IT and Services	0.118	-0.078	0.314	0.665
Banking-IT and Services	0.212	0.091	0.334	<.001
Financial Services-Hospital & Health Care	-0.057	-0.231	0.118	0.991
Marketing and Advertising-Hospital & Health Care	-0.684	-0.916	-0.452	<.001
Construction-Hospital & Health Care	-0.121	-0.378	0.135	0.895
Education Management-Hospital & Health Care	-0.113	-0.458	0.232	0.99
Computer Software-Hospital & Health Care	-0.187	-0.396	0.023	0.131
Retail-Hospital & Health Care	-0.368	-0.551	-0.185	<.001
Higher Education-Hospital & Health Care	-0.046	-0.275	0.182	1

Continued on next page

Table 1 – continued from previous page

	Difference	Left bound	Right bound	p
Banking-Hospital & Health Care	0.048	-0.121	0.217	0.997
Marketing and Advertising-Financial Services	-0.627	-0.831	-0.424	<.001
Construction-Financial Services	-0.065	-0.296	0.167	0.997
Education Management-Financial Services	-0.056	-0.383	0.27	1
Computer Software-Financial Services	-0.13	-0.308	0.048	0.379
Retail-Financial Services	-0.312	-0.457	-0.166	<.001
Higher Education-Financial Services	0.01	-0.19	0.21	1
Banking-Financial Services	0.104	-0.023	0.232	0.218
Construction-Marketing and Advertising	0.563	0.285	0.84	<.001
Education Management-Marketing and Advertising	0.571	0.21	0.932	<.001
Computer Software-Marketing and Advertising	0.497	0.263	0.732	<.001
Retail-Marketing and Advertising	0.315	0.104	0.527	<.001
Higher Education-Marketing and Advertising	0.637	0.386	0.889	<.001
Banking-Marketing and Advertising	0.732	0.532	0.931	<.001
Education Management-Construction	0.008	-0.369	0.385	1
Computer Software-Construction	-0.065	-0.324	0.194	0.999
Retail-Construction	-0.247	-0.485	-0.009	0.034
Higher Education-Construction	0.075	-0.2	0.349	0.998
Banking-Construction	0.169	-0.058	0.396	0.355
Computer Software-Education Management	-0.074	-0.42	0.273	1
Retail-Education Management	-0.256	-0.587	0.076	0.302
Higher Education-Education Management	0.066	-0.292	0.425	1
Banking-Education Management	0.161	-0.163	0.484	0.862
Retail-Computer Software	-0.182	-0.368	0.005	0.063
Higher Education-Computer Software	0.14	-0.091	0.371	0.658
Banking-Computer Software	0.234	0.062	0.407	0.001
Higher Education-Retail	0.322	0.114	0.529	<.001
Banking-Retail	0.416	0.277	0.555	<.001

Continued on next page

Table 1 – continued from previous page

	Difference	**Left bound**	**Right bound**	**p**
Banking-Higher Education	0.094	-0.101	0.289	0.88

Table 2

Descriptive statistics for all continuous features and outcome variables for both classification and regression

	Minimum	Maximum	Mean	SD
Working years	2.01	50.0	16.64	7.46
Mean job duration	.09	17.77	2.25	1.52
Last job duration (Regression)	.50	27.77	2.41	2.01
Last job duration (Classification)	0	1	.52	.50

Table 3

Frequency table of industry type of the last job

	Frequency
Information Technology and Services	19037
Hospital & Health Care	13091
Financial Services	10206
Marketing and Advertising	9710
Construction	9606
Education Management	8361
Computer Software	8038
Retail	7729
Higher Education	7203
Banking	7019

Table 4

Feature based classification models

	Accuracy		
	Industry	Years worked	Mean job duration
Logistic Regression	.53	.63	.65
Decision Tree	.53	.63	.67
Random Forest	.53	.63	.67
Naive Bayes	.51	.62	.62
Neural network	.52	.63	.66

Table 5

Feature based regression models

	R^2		
	Industry	Years worked	Mean job duration
Linear Regression	.01	.16	.22
Decision Tree	.01	.16	.22
Random Forest	.01	.16	.21
Linear SVR	.01	.15	.22
Neural network	.01	.16	.22

Table 6

The predictive accuracy of the classification algorithms

		Accuracy		
Model	Parameters	Cross-validation	Test	Training time (s)
Majority Baseline	-	.52	.50	-
Logistic Regression	penalty = "L2" solver = "liblinear"	.66	.66	.20
Decision Tree	min_samples_split: 7 max_depth: 3 min_samples_leaf' 4 max_features: 5	.65	.65	.01
Random Forest	n_estimators: 3 min_samples_split: 9 max_depth: 3 min_samples_leaf: 8 max_features: 9	.67	.66	3.85
Naive Bayes	- .52	.64 .01	.64	.01
Neural Network	activation='tanh' optimizer = "Adam" loss='categorical_crossentropy' n_hidden_layers= 2 epochs = 2	.67	.67	2.08*

Note. *Please note that the Neural Network training time strongly depends on the amount of epochs. For the classification 2 epochs were used, at that point the loss of the model stabilized.

Table 7

Correlation matrix of all classification models

	Log regression	Decision tree	Random Forest	Naive Bayes	Neural network
Logistic Regression	1				
Decision Tree	.85	1			
Random Forest	.84	.99	1		
Naive Bayes	.91	.78	.78	1	
Neural network	.93	.87	.87	.86	1

Table 8

Results of the regression task

		R^2		
Model	Parameters	Cross-validation	Test	Training time (s)
Baseline	-	0	.0	-
Linear regression	fit_intercept: True	.24	.22	.01
Decision Tree	min_samples_leaf: 6	.23	.21	.01
	min_samples_split: 8			
	max_depth: 3			
	max_features: 9			
Random Forest	n_estimators: 4	.23	.21	.03
	min_samples_split: 6			
	max_features: 6			
	min_samples_leaf: 7			
	max_depth: 5			
Linear SVR	epsilon: 0.07514395715986755	.24	.22	0.85
	loss: squared_epsilon_insensitive			
Neural Network	activation='tanh'	.25	.22	10.45*
	optimizer = "Adam"			
	loss='categorical_crossentropy'			
	n_hidden_layers= 2			
	epochs = 10			

Note. *Please note that the Neural Network training time strongly depends on the amount of epochs. For the regression 10 epochs were used, at that point the loss of the model stabilized.

Table 9

Correlation matrix of all regression models

	Linear regression	Decision tree	Random Forest	Linear SVR	Neural network
Linear Regression	1				
Decision Tree	.88	1			
Random Forest	.84	.91	1		
Linear SVR	.94	.81	.87	1	
Neural network	.98	.84	.91	.93	1

Table 10

Summary of previous research on employee turnover forecast.

Authors (year)	Data acquisition	Data horizon	Methods	Software	Economic indicator	Response variable	Estimate	Model evaluation
Bluedorn (1982)	Employee records and survey	1 year	Correlations, multiple regression	N/A	No	Number	Point with intervals	R2 = 0.22, Adjusted R2 = 0.11
Ng et al. (1991)	Survey	N/A	Hazard proportional model	BMDP 2L	No	Probability	Point with intervals	Pair t-test
Balfour and Neff (1993)	Employee records	33 months	Nonlinear logistic regression	N/A	No	Probability	Point	Chi-square values
Feeley and Barnett (1997)	Survey	60 months	Social network, logistic regression, correlation	NEGOPY, UCINET	No	Probability	Point	R2 = 0.23
Wright and Cropanzano (1998)	Survey	1 year	Hypothesis test, correlation, logistic regression	N/A	No	N/A	N/A	Correlation r = 0.34, P < 0.01
Morrow et al. (1999)	Demographic information and employee records	2 years	Logistic regression, correlation	N/A	No	Probability	Point	([Pleaseinsertintopreamble]2 log likelihood) chi-square = 193.13
Sexton et al. (2005)	Demographic information and employee records	10 years (yearly)	NN	FORTRAN	Yes	Leave or not	Point	Type I error = 0.25% Type II error = 5.83%
Hong et al. (2007)	Survey	N/A	Logit and probit model	SPSS	No	Probability	N/A	R2 = 0.5, Quadratic Probability Scores = 0.18 for training and 0.12 for test
Nagadevara et al. (2008)	Demographic information and employee records	3 years	NN, logistic regression, classification/regression trees, discriminant analysis	N/A	No	Leave or not	Point	Contingency table
Thaden et al. (2010)	Survey	2 years	Multiple regression	N/A	No	Duration	Point with intervals	R2 = 0.56, P < 0.001
Größler and Zock (2010)	Employee records	360 months	System dynamics	N/A	No	Number	Point	N/A
Saradhi and Palshikar (2011)	Survey	2 years	SVMs, random forest, Na[Pleaseinsertintopreamble]ve Bayes classifiers	N/A	No	Probability	Point	True/false positive rate and precision
Alao and Adeyemo (2013)	Employee records	28 years (yearly)	Decision tree	WEKA See5	No	Probability	Point	True/false positive rate and precision
Tews et al. (2014)	Employee records and Survey	6 months	Logistic regression	N/A	No	Probability	Point	R2 = 0.23
Collini et al. (2015)	Survey and turnover rates	1 year	Correlation and regression	N/A	No	Turnover rates	Point	No

Note. Adapted from 'Employee turnover forecasting for human resource management based on time series analysis.' , X. Zhu, W. Seaver, R. Sawhney, R. Ji, B.Holt, G. B. Sanil, and G. Upreti, 2016, Journal of Applied Statistics, p. 1-20.

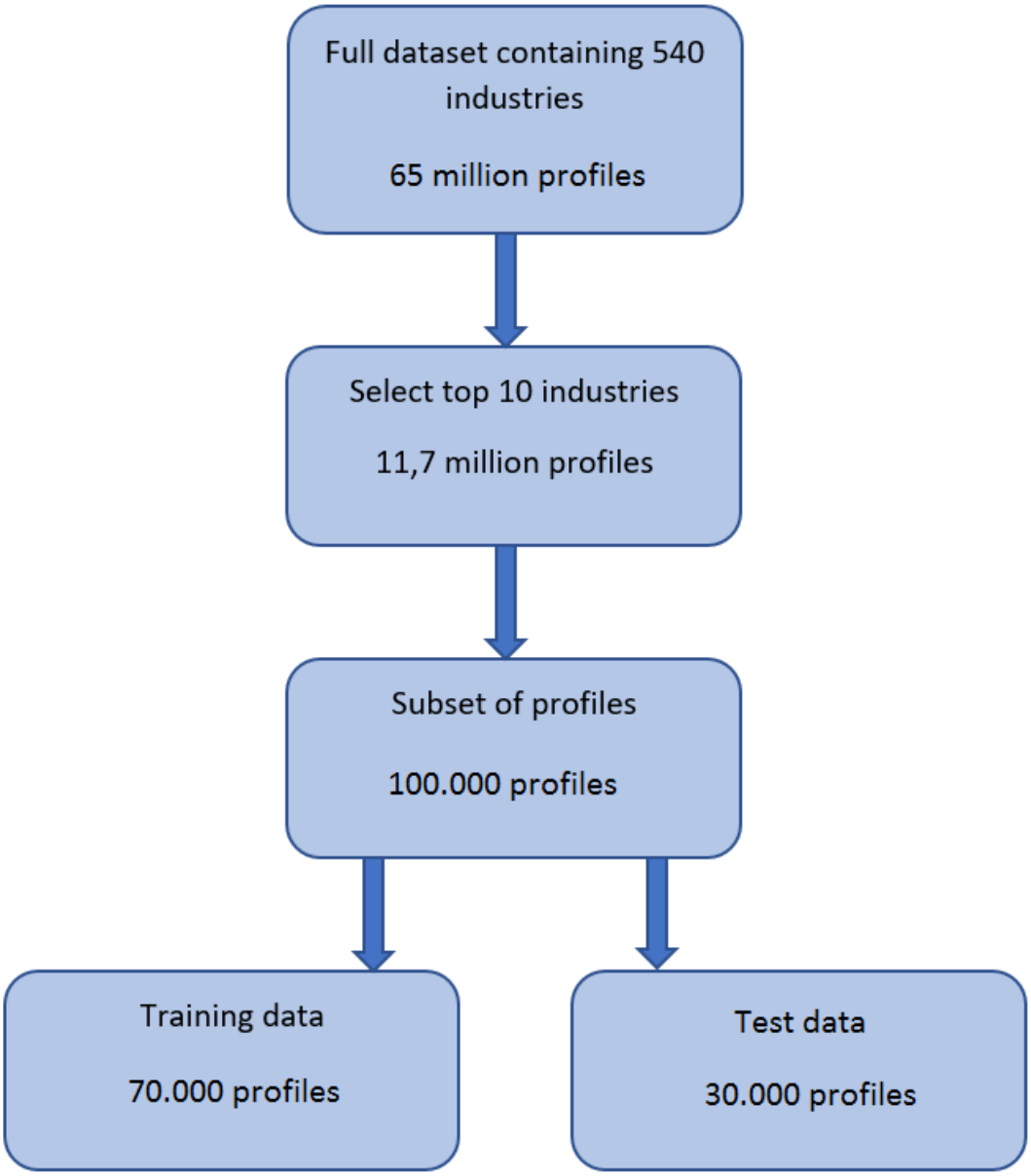

Figure 1. Flowchart describing how the subset of the data was selected

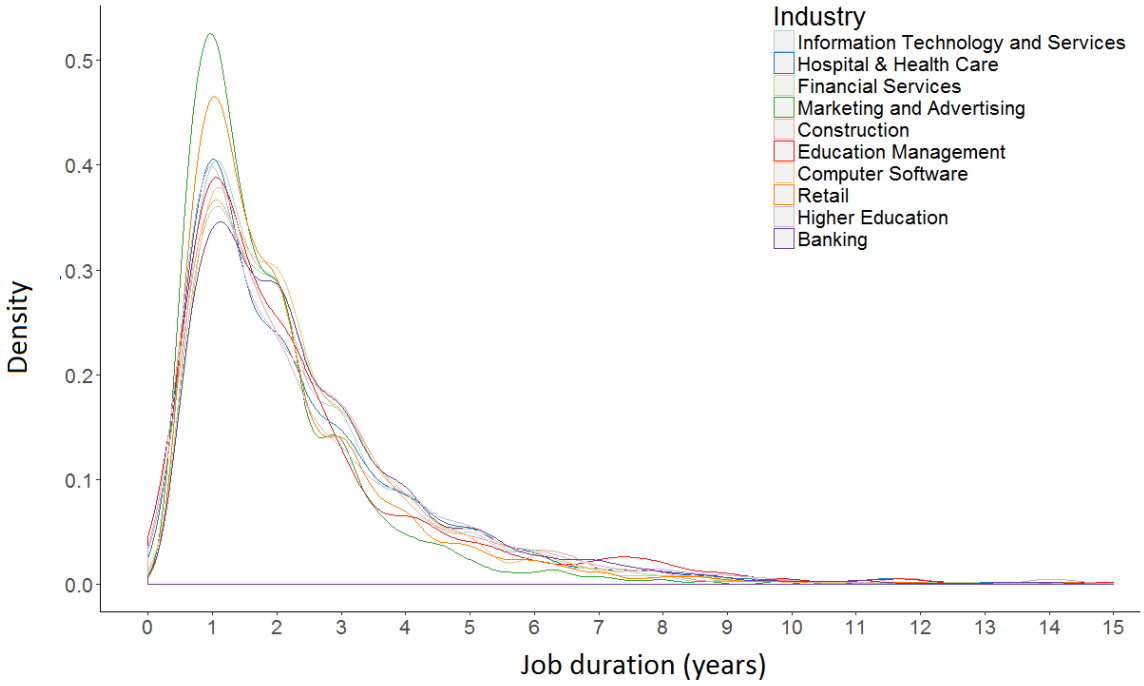

Figure 2. Density plot of the mean job duration (years) for the different industries