Predicting employee attrition with machine learning on an individual level, and the effects it could have on an organization

FREDRIK NORRMAN
Predicting employee attrition with machine learning on an individual level, and the effects it could have on an organization

by

Fredrik Norrman
Predicera uppsägninrar på en individuell nivå med machine learning, och effekterna det kan ha på en organisation

av

Fredrik Norrman
Abstract
This paper is investigating the possibility to predict employee attrition on an individual level with machine learning. The study is divided into two parts, one qualitative part which were conducted by doing interviews with selected roles where the openness to which practitioners are willing to use machine learning models to predict employee attrition, and what effects such a model could have on an organization was investigated. The second part is a quantitative part where a random forest model, support vector machine model and a logistic regression model are compared in terms of accuracy in predicting employee attrition with the usage of large human resource data sets.

Firstly, it was shown that people are willing to use machine learning models to predict employee attrition if the models were to be trusted, and if organizations that used such models were transparent in how the models were used, and to what purpose. The model comparison did not give any interesting results about the possibility to predict employee attrition with the chosen models. There were several reasons for that, where some of them were that the models were over fitted, the time of notice when a person quit was not accounted for enough and the choice of input data points. This resulted in that the accuracy could not be determined in a confident way.

Key-words
Employee attrition, Employee turnover, Machine learning, Prediction, Human resource
Sammanfattning


För det första så visades det att utövare är öppna för att använda machine learning för att predicera uppsägningar så länge modellerna som används kan litas på. Dessutom visade det sig att transparens från organisationer som använder sig av sådana modeller är viktigt, där tydighet i hur modellen används och varför och vad den används till måste kommuniceras. Av den andra delen, där de tre modellerna ställdes mot varandra, visade det sig att det var svårt att predicera uppsägningar baserat på medarbetsundersökningar. Det gick inte med säkerhet att visa att den träffsäkerheten som uppnåddes faktiskt betyder något väsentligt utan istället så visade det sig att det fanns problem som gjorde att resultaten inte blev som förväntat. Dessa problem var bland annat över fitting, uppsägningstid togs inte med i beräkningarna tillräckligt mycket, och valet av input data visade sig inte vara bra nog.

Nyckelord

Employee attrition, Employee turnover, Machine learning, Prediction, Human resource
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1 Introduction</strong></td>
<td></td>
</tr>
<tr>
<td>1.1 Background</td>
<td>4</td>
</tr>
<tr>
<td>1.2 Problematization</td>
<td>5</td>
</tr>
<tr>
<td>1.3 Purpose</td>
<td>7</td>
</tr>
<tr>
<td>1.4 Research questions</td>
<td>7</td>
</tr>
<tr>
<td>1.5 Delimitation</td>
<td>7</td>
</tr>
<tr>
<td>1.6 Expected outcomes</td>
<td>7</td>
</tr>
<tr>
<td><strong>2 Literature review</strong></td>
<td>8</td>
</tr>
<tr>
<td>2.1 Human resource management and competitive advantage</td>
<td>8</td>
</tr>
<tr>
<td>2.1.1 Employee turnover</td>
<td>8</td>
</tr>
<tr>
<td>2.1.2 Retention management</td>
<td>9</td>
</tr>
<tr>
<td>2.1.3 Factors that make people stay at their jobs</td>
<td>9</td>
</tr>
<tr>
<td>2.2 Machine learning in Human Resource Management</td>
<td>10</td>
</tr>
<tr>
<td>2.2.1 Algorithmic decision-making</td>
<td>11</td>
</tr>
<tr>
<td>2.2.2 Algorithmic Pollution</td>
<td>11</td>
</tr>
<tr>
<td>2.3 Machine learning</td>
<td>13</td>
</tr>
<tr>
<td>2.3.1 Trustworthiness of using machine learning</td>
<td>13</td>
</tr>
<tr>
<td>2.3.2 Interpretability</td>
<td>13</td>
</tr>
<tr>
<td>2.3.3 Types of machine learning</td>
<td>15</td>
</tr>
<tr>
<td>2.3.3.1 Supervised learning</td>
<td>15</td>
</tr>
<tr>
<td>2.3.3.2 Unsupervised learning</td>
<td>16</td>
</tr>
<tr>
<td>2.3.3.3 Reinforcement learning</td>
<td>16</td>
</tr>
<tr>
<td><strong>3 Method</strong></td>
<td>17</td>
</tr>
<tr>
<td>3.1 Research design</td>
<td>17</td>
</tr>
<tr>
<td>3.2 Data collection</td>
<td>18</td>
</tr>
<tr>
<td>3.2.1 Interviews</td>
<td>18</td>
</tr>
<tr>
<td>3.2.2 Quantitative data</td>
<td>19</td>
</tr>
<tr>
<td>3.2.2.1 Used datasets in experiments</td>
<td>19</td>
</tr>
<tr>
<td>3.2.3 Secondary sources</td>
<td>20</td>
</tr>
<tr>
<td>3.3 Model comparison</td>
<td>21</td>
</tr>
<tr>
<td>3.3.1 Random forest and SVM comparison</td>
<td>21</td>
</tr>
<tr>
<td>3.3.1.1 Random forest</td>
<td>22</td>
</tr>
<tr>
<td>3.3.1.2 Support vector machine</td>
<td>25</td>
</tr>
<tr>
<td>3.3.2 Multiple logistic regression model</td>
<td>27</td>
</tr>
<tr>
<td>3.3.2.1 Usage of MLR in this study</td>
<td>28</td>
</tr>
<tr>
<td>3.4 Quality of research</td>
<td>28</td>
</tr>
<tr>
<td><strong>4 Empirics</strong></td>
<td>30</td>
</tr>
<tr>
<td>4.1 Random forest and SVM results</td>
<td>30</td>
</tr>
</tbody>
</table>
Foreword
I would like to thank Populum for giving me the opportunity to write my thesis in collaboration with them. And I would also like to thank them for their willingness to help me throughout the process. Lastly I would like to thank my supervisor Gisela Bäcklander for the guidance along the making of this study. It would not have been possible without your help.
1 Introduction

Today, organizations put a lot of effort into human resource (HR) departments where one of most important tasks is to manage employee attrition to reduce employee turnover (Koys, 2006; Ajit & Punnoose, 2016). A contributing factor to that are the effects that comes with employee attrition. Replacing experienced workers who leaves for other organizations costs money in form of hiring expenses and training the new employee (Alduayj & Rajpoot, 2018; Sexton et al. 2005). Furthermore, tacit and explicit knowledge is lost when an employee leaves, and social relationships that are important can be broken (Ashworth, 2006; Droege & Hoobler, 2003; Levy, 2011).

HR employees often struggles with explaining their value creation for their organizations and one of their tasks is to make HR be more significant through better decisions (Tomassen, 2016). Today, there is an increase in HR departments trying to base decisions on data. Data-driven decisions can lead to better organizational performance which means that if HR departments can base decisions on data, they would bring value to the organization (Tomassen, 2016). The process where decisions are derived from analyzing data belongs to the field data analytics. Data analytics is used by businesses to interpret and extract information from data that can be used for decision making (Kelleher, Namee & D’Arcy, 2015). Data analytics is increasing within the field of HR and has the potential to increase the importance of HR departments for organizations (King, 2016).

A common technique for prediction purposes within data analytics is machine learning (ML) (Domingos, 2012; Kelleher, Namee & D’Arcy, 2015). Machine learning is the study of how to make computers learn from experience (Jordan & Mitchell, 2015; Ayodele, 2010; Tomassen, 2016). The idea is for an algorithm to learn from datasets, and improve when it is exposed to new data. Potentially, ML could be used in HR departments to predict employee attrition. A problem with that would be that decisions within HR departments can have big consequences for employees, their families and the organization (O’Neil, 2016; Marjanovic, Cecez-Kecmanovic & Vidgen, 2018; Deobald et al. 2019). Therefore, This paper will investigate the possibility to use machine learning to predict employee attrition. This will be done by analyzing two different strategies to predict employee attrition, and investigate people’s will to both base life-changing decisions on algorithms and to be exposed for algorithmic decisions.

This paper is organized in the following way: Firstly a short background which will introduce some areas where machine learning is used today, and the importance of managing employee turnover in organizations. Based on the introduction and background, a purpose and research questions will be presented. Secondly, a literature review on machine learning, human resource management and machine learning in human resource management theory will be presented. Thirdly, an explanation on how this paper was conducted will be presented.
Fourth, all the empirical results are presented. Fifthly, an analysis and discussion of the empirics is presented. Conclusively, a conclusion is drawn as well as potential future research.

1.1 Background

Employee turnover is playing a big part in the success of an organization, and there are several reasons for that fact (Ajit & Punnoose, 2016). As mentioned in the introduction, there are several effects on employee turnover that could have negative effects on an organization. Some of them but not all are that, firstly, employee turnover is expensive. People who voluntarily quit, needs to be replaced. A hiring process takes time, time which could be put into more productive tasks. Often, it takes time for a new employee to get up to full speed and investments in training the new employee is not an unusual thing (Koys, 2006; Alduayj & Rajpoot, 2018; Sexton et al. 2005). Secondly, the productivity can be affected. Experienced workers carry knowledge and experience which contributes to productivity. A less experienced worker is less likely to be as productive as an experienced worker and it takes time to reach the rate of productivity of a more experienced worker (Alduayj & Rajpoot, 2018; Sexton et al. 2005). Thirdly, the morale of the remaining employees tends to be affected negatively by high turnover rates. One reason for that is that when employees leave, other employees has to fill in and do tasks which were supposed to be done by someone else. The workload increases. If the employee turnover rates are high, this means a high increase in workload which can decrease motivation and morale (Ashworth, 2006; Droge & Hoobler, 2003; Levy, 2011). Fourth, employee turnover affects profit. The effects of the above affects organizational performance which in turn affects the ability to perform at the desired level. Both the costs, knowledge loss and the reduced productivity therefore also affects the profit.

The above effects of employee turnover are reasons why it is essential that organizations use effective and intelligent strategies to lower employee turnover. Being able to predict employee turnover would add tools to the toolbox when building such strategies. These strategies would increase the chances of preventing employee turnover, and therefore also minimize the risk of getting all the negative consequences that was listed above (Ajit & Punnoose, 2016; Sexton et al. 2005).

Today, organizations try to predict employee turnover and use that information to be able to lower the employee turnover rates. However, the majority of predictions on employee turnover are today focused on turnover on an organization as a whole and not on an individual level. Predicting employee turnover on a general level in an organization could help with identifying that turnover is a problem and could give life to a retention strategy and to an analysis of why people are leaving or not. However, predicting employee turnover on an individual level could give organizations even better chances of working proactively and take necessary action to minimize employee attrition and therefore have a positive effect on
employee turnover. That in turn could lead to that the organization can keep expertise knowledge within the company, keep and improve productivity, maintain social sustainability and workplace morale and prevent disruptions within projects, which would have a positive effect on the organizational performance and therefore also the competitive advantage. (Ajit & Punnoose, 2016; Tomassen, 2016; Marjanovic, Cecez-Kecmanovic & Vidgen, 2018)

The models that have been used are often traditional prediction models such as the logistic regression model and the results are often inconsistent (Yilnaz & Kaynar, 2011; Paliwal & Kumar, 2009). In the last 20 years, the traditional models have been challenged by machine learning models for prediction purposes in several fields (Domingos, 2012; Paliwal & Kumar, 2009). This is because the increased accuracy and the ability to solve more complex problems that often comes with a machine learning model (Lundberg & Su-In, 2017; Paliwal & Kumar, 2009). And so could it be even in HR with predicting employee turnover.

There are just a few studies made on predicting employee turnover, or employee attrition rather, on an individual level. That is, trying to predict when an employee leaves. Sexton et al. (2005) made a study where they tried to predict employee attrition on individuals with the use of machine learning. The results were somewhat successful, however Sexton et al. (2005) argues that further research and several similar studies needs to be conducted before any conclusion that employee turnover can be predicted on an individual level can be drawn. A similar study was made by Ajit & Punnoose (2016). Their results were also promising in predicting employee turnover but they suggest just as Sexton et al (2005) that further research and testing in several organizations has to be made before a conclusion about a suitable model that can be used to predict employee turnover. Furthermore both Sexton et al. (2005) and Ajit & Punnoose (2016) identifies several potential organizational benefits if a model would be successful in predicting employee turnover.

The trend over the last decade has been that leaders are growing an interest in using machine learning within their organizations for decision making (Alduayj & Rajpoot, 2018). One reason for that is that researchers have found that there exists positive effects on the organizational performance when decisions are based on data, and machine learning is widely used in predictive data analytics to create models which base decisions on learning from large data sets (Tomassen, 2016; Kelleher, Namee & D’Arcy, 2015). Using machine learning to predict employee attrition on an individual level have the potential to improve retention strategies that could lower employee turnover and therefore also have positive effects on an organization. This would also increase the value creation for HR departments which would increase their significance for organizations.

Since the use of machine learning to predict employee attrition affect individuals, and could be seen as sensitive information, it is necessary that people would be open to use such models. Ribeiro, Singh & Guestrin (2016) argues that individuals need to trust machine
learning models to use them, regardless of the results. Therefore it is also necessary to investigate people's will to be exposed to such models.

1.2 Problematization

Managing employee attrition to obtain low and healthy turnover rates in an organization is of great importance to maintain organizational performance and therefore also the competitive advantage (Shaw, 2010; Ajit & Punnoose, 2016; Koys, 2006). The pressure on HR departments to provide value to the organization has led to the introduction of data-driven decisions and machine learning (Tomassen, 2016). Today, high employee attrition rates are considered as an issue for organizations and that has in turn increased the responsibility on HR departments to manage employee attrition rates at a level that is healthy (Park & Shaw, 2013; Glebbeek & Bax, 2017; Shaw, 2010).

1.3 Purpose

The purpose of this study is to investigate how a machine learning model can be applied to large human resource (HR) data sets to be able to predict employee attrition and to investigate practitioners openness to use such models for decision making within an organization. That is, this study will aim to analyze the possibility to predict employee attrition based on personnel HR data sets and what effects practitioners think that being able to predict employee attrition could have on organizational performance.

1.4 Research questions

The research question used in this study are divided into two main questions 1 and 2 and one sub question 2b and are as follows:

1. How does predictions with a machine learning model perform compared to a simpler regression analysis model when applied to HR data sets to predict employee attrition?

2. How does understanding a ML model, that predicts employee attrition, affect people’s will to use them?

2b. What do practitioners believe that such a model can contribute with to an organization?

1.5 Delimitation

This study only investigates the possibility to predict the phenomenon of employee attrition and not other employee happenings, and what perceptions and beliefs practitioners think that such a prediction could have on organizational performance. The models that are used are limited to two machine models because of the large amount of available models and the time frame for this project and the time it takes to get acquainted with different ML methods. The
qualitative part of the study is limited to Sweden, and only employees in Sweden are interviewed.

1.6 Expected outcomes
This study contributes to knowledge by investigating the possibility to apply machine learning to human resource data to be able to get a deeper understanding of employee attrition and organizational performance and to predict future employee attrition. The study also strengthen knowledge about how practitioners perceive that employee attrition could affect the performance of an organization and therefore also the competitive advantage. This could give many interesting results. For example if an organization would be able to predict employee attrition, actions could be made to prevent it or to minimize the damage it could make.

2 Literature review
In this section all of the necessary literature for the purpose of this paper is presented. This includes a description of how human resource management can have an impact on organizations, the usage of machine learning in human resources and an introduction to what machine learning is and what it is usually used for.

2.1 Human resource management and competitive advantage
A goal which all organizations have is to gain a competitive advantage against their competition. One important area to address to gain competitive advantage is the management of human resources (Human resource management, HRM). It has been shown that the level of happiness and satisfaction an employee have will affect their engagement and how well they will perform, and in turn have an impact on the organizational performance (Anitha, 2014). Effective HRM has the potential to have a positive impact on employees which is likely to increase their performance and thus increase an organization’s competitive advantage (Koys, 2006).

2.1.1 Employee turnover
A big part of HRM is to keep employees satisfied. A study by Ryan, Schmit and Johnson (1996), demonstrated that employee satisfaction was linked to employee turnover. A higher satisfaction tends to lower the employee turnover. Several empirical studies have shown that employee turnover itself can, and often has, an impact on organizational effectiveness which affects competitive advantage (Koys, 2006; Davis, 2013; Morrow & McElroy, 2007; Dess & Shaw, 2001). Some ways in which employee turnover affects organizational effectiveness are: (1) when an employee leaves their tacit and explicit knowledge is lost which decreases
the overall knowledge within that particular organization, the productivity of the organization could be affected in a negative way, (2) increased workload on other employees, (3) have a negative impact on company morale, (4) lower turnover rates means that there will be less hiring which will lower costs such as activities for training and employment process (Koys, 2006; James & Mathew, 2012). There are both tangible and intangible effects of employee turnover. And to get a deep understanding of the actual effects and how employee turnover affects the competitive advantage, it is important to consider them both.

Kagmar et al. (2006) argues that some employee turnover can be beneficial for an organization but that turnover in most cases are very costly and can disrupt the workflow. As an example how costly employee turnover could be we turn our eyes towards the U.S. fast food industry. This is an industry which is exposed to high employee turnover, and only retraining costs of new employees per year is as high as $4.3 billion (Kagmar et al, 2006). The first step of being able to reduce employee turnover is to understand why the phenomenon is happening in the first place, and that is extremely important. This is because, knowing what causes the problem, gives organizations the chance to take action.

2.1.2 Retention management

Retention management is a part of HRM and is the study of how organizations could work to keep their employees for a maximum amount of time. To be able to obtain effective retention management, organizations must acknowledge and analyze the turnover situation (James & Mathew, 2012). The knowledge about how the turnover situation is not enough. In order for organizations to act, they also need a strategic retention plan with concrete actions to undertake in order to maintain their employees and studies have shown that organizations with retention strategies in place increases the chances of retaining their employees and that retaining employees will create attractiveness which can attract new employees as well (James & Mathew, 2012). Furthermore, James & Mathew (2012) argue that it is the responsibility of the organization to actively work to keep their best employees and that a failure in that regard will lead to that their best performing employees will be lost. This is why retention management has evolved to an important part of HRM and why the effects could have such a big impact on competitive advantage.

2.1.3 Factors that makes people stay at their jobs

Changing organizations voluntary can be fulfilling for an individual, but can also be tough and demanding (Mitchell et al. 2003). Several researchers over the years have looked into why employees leave their organizations for another (Mitchell et al. 2003; Davis, 2013; Ladelsky & Catană, 2013). The general answer is very simple. If an employee is not satisfied with their job and does not have a particular commitment to the organization, that employee tend to leave (Mitchell et al. 2003).
There are several factors that affect people’s will to stay in their organizations, both external and internal. External factors could be job opportunities or people moving to other places (Mitchell et al. 2003; Hausknecht, Rodda & Howard, 2009). Internal factors are embedded within job satisfaction and organizational commitment (Mitchell et al. 2003). Breaking down job satisfaction into several smaller parts results in detailed factors that play a part in why employees choose to stay in their organizations. Factors that affect job satisfaction are that employees need to feel connected with other employees and activities within the organization, employees need to feel that they belong in their organization, employees need to have the feeling that they have their organizations support to do their task and that employees feel that their workload is reasonable (Mitchell et al. 2003; Hausknecht, Rodda & Howard, 2009; Cho, Johanson & Guchait, 2008).

2.2 Machine learning in Human Resource Management

The time it has taken for machine learning to enter the realm of human resources has been long and it hasn't been used until just a few years back (Faggella, 2019). According to Linkedin (2018) only 22% of organizations have initiated analytics into their human resource departments and it is uncertain on how well these analytics are implemented. Fields such as marketing has a rich access to big data, take for example the sales of a product in a country. There will be data such as how many products has been sold, when they have been sold and how many observations the product has gotten. With those conditions, ML is a very effective tool to use to analyze big sets of data. (Cappelli, Tambe & Yakubovic, 2018)

The use of machine learning in HR presents several problems and challenges that has to be accounted for. First of all, the data stored in human resources are very small in contrast to what data-sets used in data science are, and often data is not stored at all (Cappelli, Tambe & Yakubovic, 2018). Even a big organization with 10000 employees are nowhere close to contain as much HR data as product sales data, for example. Secondly, the decisions within human resource management could get big consequences (Sutherland & Wocke, 2011; Cappelli, Tambe & Yakubovic, 2018). For example, a decision of whom gets fired and who gets hired. Thirdly, measuring the performance of individuals are often hard due to that complex roles often depend on other roles - teamwork - and individual performance can be hard to break down from group performance (Cappelli, Tambe & Yakubovic, 2018; Murray & Enarson, 2007). Fourth, deciding whether or not a person should be hired is not only based on tangible qualities but also on social and psychological relationships between employees (Kluger et al. 2002; Cappelli, Tambe & Yakubovic, 2018). Lastly, a machine learning algorithm needs to be trained and the algorithm will most likely be characterised by that. This could, for example be a problem when hiring new employees if a machine learning algorithm is trained on the current workforce and say that the majority of the workforce are white men, the algorithm could be biased towards white men when looking for new candidates (Cappelli, Tambe & Yakubovic, 2018). This exact example happened Amazon in 2018 and they
specifically had removed gender from their model as a criterion, but even then the model learned in a way so that it was biased towards one specific group of people.

However, the potential for machine learning within HR is huge. That is due to that HR operations and outcomes affect organizational performance in different ways. These are operations such as: onboarding for new employees, training new and old employees, identify good and bad performance, determine who should be promoted, employee retention and employee benefits. (Cappelli, Tambe & Yakubovic, 2018)

2.2.1 Algorithmic decision-making

It is becoming more common that algorithms are used for decision making in both people’s private and business lives (Marjanovic, Cecez-Kecmanovic & Vidgen, 2018). These decisions could play a big role and often they could even be life-changing. In HR it could be screening job applicants in the hiring process, deciding who gets to move on in the process or not, and it could also be decisions about job performance (Marjanovic, Cecez-Kecmanovic & Vidgen, 2018; Deobald et al. 2019). While algorithmic decision making is beneficial in many cases, letting an algorithm make decisions for us could have negative impacts (Marjanovic, Cecez-Kecmanovic & Vidgen, 2018; Deobald et al. 2019). Taking the hiring process as an example where algorithms can screen thousands of job applications in just a blink of an eye, deciding which candidates are interesting, compared to the time it would take for a human to go through all the applications manually it is clear that using an algorithm would save a lot of time. In the same exact example, just imagine the case where the best candidates are automatically removed because of the algorithm.

In transformative services, that is services that transforms or changes people’s lives, the consequences of algorithmic decision making could potentially be disastrous and hurt both individuals and their families but also organizations (O’Neil, 2016; Marjanovic, Cecez-Kecmanovic & Vidgen, 2018). Transformative services are usually services such as healthcare, schools and elderly cared. However, with the definition that a transformative service is a service that changes people’s lives, HR decisions can be included.

2.2.2 Algorithmic Pollution

Algorithmic pollution is an expression that Marjanovic, Cecez-Kecmanovic & Vidgen (2018) introduced as a label on the negative consequences that comes from algorithmic decision-making. Algorithmic pollution is a new type of externality. In the same sense as environmental pollution, but in a social setting where people lives, for example in our everyday life at work or at home. Algorithmic pollution is not regulated or even recognized in many cases, unlike environmental pollution which have clear regulations and is globally recognized as a massive problem (Marjanovic, Cecez-Kecmanovic & Vidgen, 2018;
Jayaswal, Sahu & Gurjar, 2017). The general opinion about algorithms that they are often seen as neutral, objective, efficient and that they can produce better decisions than humans has hindered the development of regularization of algorithms use and the pollution that they leave behind (Marjanovic, Cecez-Kecmanovic & Vidgen, 2018).

More specifically Marjanovic, Cecez-Kecmanovic & Vidgen (2018) argue that algorithmic pollution are the negative consequences that people and organizations get from using algorithms when these consequences cannot, in an effective way, be detected and erased so that the consequences have a negative impact on the social environment. More specifically they define algorithmic pollution as follows: “Algorithmic pollution denotes the presence of unjustified, unfair, discriminatory, or other harmful consequences of algorithmic decision-making for individuals, groups, organizations, sections of the population, the economy, or society at large”.

According to Marjanovic, Cecez-Kecmanovic & Vidgen (2018) algorithmic pollution should be viewed similarly to how environmental pollution is viewed. That in the same way that people deserve to live in an environment that is pollution-free in a way so that the health isn’t jeopardized, people deserve to live in an environment where algorithmic pollution doesn’t have a negative impact on people’s health as well. The importance of this is growing as the use of decision-making algorithms are becoming the new normal (Marjanovic, Cecez-Kecmanovic & Vidgen, 2018; Hall & Gill, 2019). Some degree of pollution will always occur and can be acceptable, the problem arises when the algorithmic pollution creates systematic inequality between individuals or has negative effects on the society (Marjanovic, Cecez-Kecmanovic & Vidgen, 2018). That is the challenge that needs to be solved, to find these consequences so that they can be dealt with.

A problem which is emerging with finding the consequences of algorithmic decision-making is that how these algorithms actually work and interact is something that people who are affected by these decisions can’t find out, and in many situations they cant even know that there is an algorithm that has made a decision that has put them in a specific situation (Kitchin, 2017; Marjanovic, Cecez-Kecmanovic & Vidgen, 2018). And the problems with how the algorithms work and interact goes even further. How these algorithms work is not only hidden for the people who are affected by the consequences, but in many cases also hidden from the people who has made the algorithm as well (Kitchin, 2017; Marjanovic, Cecez-Kecmanovic & Vidgen, 2018). And that intimidating fact bring us into interpretability which will be talked about in the next section. One important part of interpretability is to be able to know when a machine learning decision is wrongly made (Hall & Gill, 2019). And decisions that are wrongly made in algorithmic decision-making has a high chance of contributing to algorithmic pollution.
2.3 Machine learning

Machine learning is the study of how to make computers learn from experience (Jordan & Mitchell, 2015; Ayodele, 2010; Tomassen, 2016). That is, to teach the computer to predict outcomes, based on examples. The data which a machine learning algorithm uses plays a big part of its own success. The possibility to solve bigger and more complex problems grow when the amount of available data grow (Domingos, 2012). Machine learning techniques are widely used today, and it can be found in areas such as in cars, the stock market or health care (Domingos, 2012). However, the usage of ML models are limited to people’s will to use them (Ribeiro, Singh & Guestrin, 2016). There are several types of machine learning techniques available: supervised, unsupervised, and reinforcement learning (Ayodele, 2010). The goal is the same but the approach and what prerequisites are to be fulfilled are different.

2.3.1 Trustworthiness of using machine learning

Machine learning models and predictions has been shown to accomplish things that humans cannot. However, it does not matter how good the results of a machine learning approach are if the users can’t trust them. Ribeiro, Singh & Guestrin (2016) mentioned “if the users do not trust a model or a prediction, they will not use it.” and that there are two definitions of trust: (1) “trusting a prediction”, that is a user needs to trust a prediction enough so that the knowledge gained from that prediction is used, (2) “trusting a model”, that is a user needs to trust a model to behave in expected fashion if it is used. The two are related to each other, but the first one refers to one specific prediction and the second one refers to trusting a model before it is even used in real world situations. Both of these are rooted in how much the user understands the behaviour of the model so that it does not only appear as a magical box where the middle step between the input and the output is unintelligible (Ribeiro, Singh & Guestrin, 2016). Trust in machine learning models becomes even more important when the outcomes are supposed to be used in decision making. And to understand what the model actually does plays a big part. Take for example in an area such as healthcare where ML is used for several different tasks such as medical diagnosis. To be able to act upon the result from a ML diagnose, there has to be trust in the model used since the consequences could be catastrophic. So, the usage of machine learning lies in the humans trust in different models and for us to be able to use machine learning in areas such as healthcare or autonomous vehicles we have to understand the model and know when they are wrong (Heaven, 2020).

2.3.2 Interpretability

As mentioned, for a user to trust a machine learning model it needs to be understandable. That leads us to interpretability. An explanation of how a model works has to be
interpretable, that is, the explanation has to feed the user qualitative information on how the input and the outcome are interrelated in a way that is understandable by humans (Hall & Gill, 2019; Ribeiro, Singh & Guestrin, 2016). However interpretability in machine learning is subjective and what is interpretable to some users, may not be to others (Molnar, 2020).

The need to be able to interpret and understand machine learning becomes increasingly important for people who are put in a situation where a machine learning algorithm is making a decision for them (Hall & Gill, 2019; Ribeiro, Singh & Guestrin, 2016; Domingos, 2012). There are two important drivers to why we need to be able to understand the machine learning that is used: (1) so that humans can learn from machine learning and (2) to be able to know when a machine learning decision is wrongly made. Imagine being denied a loan on a bank based on a machine learning algorithm. Without knowing how the decision was made it would be almost impossible to argue for your cause. Hall & Gill (2019) argue that the ability to explain in detail how a machine learning algorithm is making its decision is central to achieve interpretable machine learning.

The common person's ability to interpret machine learning has always been lagging behind the technique itself, and even machine learning professionals have a hard time knowing everything that is going on inside machine learning algorithms. That is due to the rapid evolvement of machine learning. But it is also the rapid evolvement of machine learning and the increased usage in our day-to-day lives that will most likely give consumers, which is almost everyone, the motivation to increase their understanding for machine learning. For example how Spotify can make personalized playlists or how the advertisement on Facebook is personalized. This will also put more pressure on machine learning professionals to improve their ability to share their understanding with others. (Hall & Gill, 2019)

Interpretability can be divided into two parts: local interpretability and global interpretability. Local interpretability is targeted to understand the relationship between the predictions made by a model and the input variables in small subsets (Hall & Gill, 2019). For example, smaller clusters of the prediction or single row data and the predicted outcome. Global interpretability refers to understanding the relationship between the predicted outcomes and their input data on larger segments of data (Hall & Gill, 2019). The global interpretability helps with understanding the relation between inputs and the predictions, but looking at a model as a whole can often give an approximation of how well the model actually works. Local interpretability lets the used look at smaller parts of the prediction and could be more descriptive of how well a model is performing (Hall & Gill, 2019). To get the best understanding of how well a model is performing, both local and global interpretability should be considered.
2.3.3 Types of machine learning

There are several types of machine learning techniques available: supervised, unsupervised, and reinforcement learning (Ayodele, 2010; Libbrecht & Noble, 2015). The goal is the same but the approach and what prerequisites are to be fulfilled are different.

2.3.3.1 Supervised learning

A supervised machine learning algorithm are given examples to train on, consisting of both input data and output data (Ayodele, 2010; Molnar, 2020). The algorithm is trained by connecting links between the input data and given output data. In the training, input data has corresponding output data which enables the algorithm to see patterns between input data and output data, so that the algorithm can learn what input data that is linked with correct output data (Ayodele, 2010). The goal is divided into two parts. The first is that the algorithm learns to connect input data to correct output data (i.e. classifications). The second, which is the ultimate goal, is for the algorithm to be able to be given a new set if input data and with that predict accurate outcomes. That is, in the learning stages the algorithm is given the "correct answers" to learn the mapping between the two. In the usage stage the algorithm is given new input data and, based on the experience, it then predicts outputs. The accuracy level of supervised machine learning algorithms is determined by using a test data set with both inputs and outputs, give the inputs to the trained algorithm and verify by the corresponding outputs within the test data set. For example if we have 100 inputs and 100 outputs, if 95 of those inputs are classified to the correct output we could say that the accuracy of the algorithm is 95%. The accuracy of an algorithm can be increased (or decreased) by altering the amount of training.

It is important that the training of a supervised machine learning algorithm is as general for the present problem as possible (Domingos, 2012). The reason for this is that when the algorithm is used in a real-world problem, it is extremely unlikely that it has been trained on exactly identical examples. So it is necessary that the algorithm is generalized in a way so that it can solve new problems. For that reason, it is pivotal that when an algorithm is tested, it should be tested with real-world data and not tested towards the data which the algorithm used to train (Domingos, 2012).

There are two types of supervised learning algorithms: regression and classification (Domingos, 2012). Regression is used when the problem is to predict a continuous output, that is a value of some sort (Ayodele, 2010). Classification is used when the problem is to find a discrete value to classify data (Kotsiantis, 2007; Ayodele, 2010). As an example we will look at prediction of rain on a cloudy day. A classification algorithm would be used to predict whether it will rain or not - yes or no - and a regression algorithm would be used to predict how much it would rain, that is the amount of rain.
2.3.3.2 Unsupervised learning

In contrast to supervised learning, an unsupervised learning algorithm does not take any outputs into account. That is, the algorithm is learning patterns from the data only from example inputs without given output (Gentleman et al. 2008). The task here is instead to describe how the input is clustered (Ayodele, 2010). Clustering is a way of organizing data which belongs together in some way in different groups (Gentleman et al. 2008). As an example we could look at splitting fruits into groups. Consider a basket with apples, pears and oranges. The task for the algorithm is to divide the fruits into separate groups with the same fruits. An unsupervised learning algorithm can without knowing the outputs see similarities between all the apples and arrange them separately into one group, in the same way it can see similarities between all the oranges and order them into one group etc. until the basket is empty and the fruits have been divided into groups, see figure 2.1 below.

![Unsupervised ML algorithm](image)

*Figure 2.1. A simple example of how an unsupervised ML algorithm can divide fruits from a fruit basket.*

2.3.3.3 Reinforcement learning

Reinforcement learning is a mix of both supervised and unsupervised learning. An algorithm with this approach is trained with given inputs without correct outputs. However the solution that the algorithm suggests comes with positive or negative feedback (Ayodele, 2010). That is, if the algorithm’s suggested solution is correct then a positive feedback is received and if the suggested solution is wrong a negative feedback is received.
3 Method

In this section the choices of research approach and why these choices fit this study is presented. The following subsections include choice of research design, how the data that is used within this paper was collected, how the chosen models were compared to each other and at last why the research that has been conducted is of high quality.

3.1 Research design

This study was conducted with both a qualitative and quantitative study approach. The work was divided into two different stages. The first stage was the quantitative part, which were consisting of a comparison of two different models, one regression analysis model and two machine learning models, to predict employee attrition based on employee experience. All the data that was used to train and compare the models was collected from approximately 35 000 pulse surveys. Pulse surveys are short surveys that employees within an organization answers. They measure several values and are adjustable so that different organizations can use tailor-made surveys for their purpose. A quantitative approach was well suited for this part of the study since a quantitative approach is often used when mathematical methods are used to analyze collected numerical data, and specifically in statistics (Creswell, 2009).

In the second part, which was the qualitative part, interviews were made to gather tangible and relevant data to see how open and willing people are to use machine learning models to predict employee attrition, and to see how these people believe that such a model could contribute to an organization. A qualitative approach was well suited for the second part of the study since the problem was to explore new theories and ideas, and in depth understand what people actually think about using ML models (Babor et al. 2017; Shah & Corley, 2006). A qualitative study, as in the second stage of this study, allows for the researcher to study a phenomenon using several different data sources which ensure that the research problem is investigated from several different angles (Baxter & Jack, 2008). Babor et al. (2017), Baxter & Jack (2008) and Shah & Corley (2006) also argues that a qualitative study decrease the impact of the researcher’s personal opinions (Babor et al. 2017; Shah & Corley, 2006).

The study started with an exploratory and inductive research approach, and throughout the process the approach jumped back and forth between an inductive and deductive setting hence an abductive approach was used. It was decided that this was fitting this study well because of the small amount of research that had been done in predicting employee attrition on an individual level. And there was already a large amount of research available on how organizations are affected by employee turnover. An exploratory approach gives the researcher the ability to change direction when and if needed to. And as in the case was for
this study, it is a good approach to use when it does not exist a large amount of research within the field (Saunders, Lewis & Thornhill, 2012).

Both primary and secondary sources were used in the study. Primary sources were received through data sets acquired by a collaborating company, which were used in the comparison of the prediction models, and interviews. Secondary sources include earlier research such as scientific articles, academic books and reviews.

3.2 Data collection

3.2.1 Interviews
One of the primary sources used in this study were interviews. 10 interviews were made. The interviews were conducted with employees within human resource departments, line managers and employees that work beneath line managers. These type of roles were chosen because they would all be exposed to the usage of machine learning models when trying to predict employee attrition. And because of the potentially broad use of such models, it was important to gather relevant data from several different roles which would have contact with the models in some way. The interviews were conducted in 8 different organizations. This was to get a more general result which could be applied to a wider span of organizations instead of just one or a few organizations. All interviewees were informed that the data collected from interviews were anonymous before the interview started. This increased the reliability in the answers and lowered bias, and increased the chances that the interviewees would not twist the answers to their benefit (Shah & Corley, 2006).

All of the interviews were face-to-face or through a video conference call and all of the interviews was recorded with a recording device. Each interviewee was asked before the recording device was used. The initial plan was to only have face-to-face interviews, but due to the coronavirus (covid-19) it was not possible in some cases. The interview length were approximately 20-30 minute. All of the interviews were made in swedish and the transcriptions were in swedish as well. However since this paper is written in english, all quotes that are present within the paper were translated to english. From the transcripts, the most common and interesting answers in relation to the research questions were then presented in the empirics section.

In order to get a deeper understanding, and to increase the credibility and involvement of the interviewee, a semi-structured interview approach was used (Miles & Gilbert, 2005; Creswell, 2009). The semi-structured approach gave each interview flexibility which allowed them to be different from each other. And this, according to Miles & Gilbert (2005) and Blomkvist & Hallin (2015), could generate answers that are more interesting for the result
than the researcher initially thought of. This approach also minimizes the risk that the researcher’s opinion would have an impact on the result and therefore lowers the risk of bias.

Down below follows a table with interviewees and which roles they had. Due to anonymity, all the names have been substituted to letters A-J.

<table>
<thead>
<tr>
<th>Interviewee</th>
<th>Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Team leader</td>
</tr>
<tr>
<td>B</td>
<td>Business analyst</td>
</tr>
<tr>
<td>C</td>
<td>Salesman</td>
</tr>
<tr>
<td>D</td>
<td>HR Specialist</td>
</tr>
<tr>
<td>E</td>
<td>Team leader</td>
</tr>
<tr>
<td>F</td>
<td>HR worker</td>
</tr>
<tr>
<td>G</td>
<td>Product owner</td>
</tr>
<tr>
<td>H</td>
<td>Manager</td>
</tr>
<tr>
<td>I</td>
<td>HR worker</td>
</tr>
<tr>
<td>J</td>
<td>HR Specialist</td>
</tr>
</tbody>
</table>

*Table 3.1: Overview of input variables and rows for each data set.*

### 3.2.2 Quantitative data

Another primary source, apart from interviews, were the data acquired from pulse-surveys that was used to train and test the models. All of the survey answers were collected from a collaborating company. The company where the surveys were conducted was a company within the welfare business and had 26000 employees. The values that was measured in the surveys were clarity, value, efficiency, workload, community and enthusiasm. The surveys were conducted in Sweden and therefore only contained data about employees in Sweden.

#### 3.2.2.1 Used datasets in experiments

Two different sets of input variables were used, one with 6 and one with 12 variables. Each dataset also had one output variable per data entry (one for each row), namely if the person had quit their job or not before the next pulse survey. For dataset 1-3, the input variables was clarity, value, efficiency, workload, community and enthusiasm.

- Clarity is a measure on how clear an employee think that their organisations plan is.
- Value is a measure on how well an employee think that they spend time on doing the right things.
- Efficiency is a measure on how well an employee think that the prerequisites to do their job are.
- Workload is a measure on how employees think that they have time to do their job.
- Community is a measure on how well an employee feel that they belong at work.
- Enthusiasm is a measure on how happy an employee is about their job.

These variables were chosen because it was decided that they all affect the factor job satisfaction which was presented in section 2.1.3, and that job satisfaction plays a major part in employees will to change organizations.

All these variables was also used in dataset 4-5, with the addition of the change in clarity, value, efficiency, workload, community and enthusiasm between two different pulse surveys. The time between these two pulse surveys were 2 months. This was made to add time perspective, since a decision such as quitting a job is often something that has been growing over time. Each variable can take on 4 different values, 0-3. The higher number, the better the score. The output variable is binary, which means that it could only be of value 0 or 1, whether a person quit their job or not. Where 1 is meaning that the person did quit the job.

In table 3.2 below, a representation of the datasets that was used for both the random forest model, the support vector machine model and the multiple logistic regression model can be found.

<table>
<thead>
<tr>
<th># Data set</th>
<th># Input variables for each row</th>
<th>Total amount of input rows</th>
<th># persons who did not quit their job</th>
<th># persons who quit their job</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>17124</td>
<td>16640</td>
<td>484</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>11331</td>
<td>10000</td>
<td>1331</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>5204</td>
<td>3873</td>
<td>1331</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>17124</td>
<td>16638</td>
<td>486</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td>5106</td>
<td>4620</td>
<td>486</td>
</tr>
</tbody>
</table>

Table 3.2: Overview of input variables and rows for each data set.

3.2.3 Secondary sources

Secondary sources was acquired through databases such as google scholar, web of science and the KTH Primo. The main keywords that were used was “Human resource management”, “Artificial Intelligence”, “Machine learning”, “Random forest”, “Support vector machine”, “Regression analysis”, “Multiple regression” and “Prediction models”. These keywords led to many insightful publications and researchers, where some of the researchers was found to have several interesting publications within the subject of the study and these were also used.
3.3 Model comparison

For the experiment in this study, three different models were used. One multiple logistic regression model and two machine learning models, namely random forest and support vector machine. The random forest model was chosen because it is one of the most used models for classification problems and is known for accurate predictions and can handle a variety of sizes of datasets (Yiu, 2019; Belgiu & Dragut, 2016; Hastie, Tibshirani & Friedman, 2008; Oshiro, Perez & Baranauskas, 2012). The support vector machine model was chosen because of the ability to work for several different problems, both for regression and classification, linear and non-linear problems (Pupale, 2018). The reason why two machine learning models were used was so that the results from the random forest model could be compared to another machine learning model and not only the multiple logistic regression model.

The machine learning models and the logistic regression model were used in order to put them against each other to analyze performance, simplicity and interpretability. The reason for these two specific strategies were that regression models are common to use in predictions (Knofczynski & Mundfrom, 2007). And the machine learning models was used to see if they could fit as a prediction model when trying to predict employee events, in this case employee attrition, with better accuracy and to see if interpretability and understanding of these models mattered.

3.3.1 Random forest and SVM comparison

As mentioned above, the two different machine learning models that were used was random forest and support vector machine. The programming language R (version R 3.6.3 with GUI version 1.70) was used to test both models. R was decided to be suitable since it is created for statistical computing and used specifically for techniques such as linear and nonlinear modeling, regression analysis, classification analysis, time series analysis, clustering etc. (The R project, 2020). Within R, the package called “Caret” (Classification and Regression Training) was used. The package was used because it fulfilled the goal of bringing functions that could be used for predictive machine learning models in a user friendly way, where integrated random forest and support vector machine functions were included (Kuhn, 2019).

Both models were tested five times each. Five input data sets were used and each model was tested once for each input data set which resulted in 10 tests in total. Number of input variables, number of rows, and how the data sets were partitioned with employees who quit their job and who did not quit their job can be seen in table 3.1 above. The amount of tests was decided for two different reasons: (1) The first input data set did not give enough interesting results, so to minimize the risk of corrupt data and faulty data several different data sets were used, and (2) the results from all tests, after five tests for each model, yielded similar results and it was then decided that changes in input data that was requiring
reasonable work effort would not change the results of the chosen models. Furthermore, the
data sets that was used was split into two subsets of 20 and 80 percent of the input data set.
This was so that the models were trained on 80 percent of the input data, and the remaining
20 percent was used as a validation set to test the accuracy of the models. The results from
each model were then presented in the empirics section.

3.3.1.1 Random forest

The random forest model is a combination of several decision trees that work together. An
overview of a decision tree can be seen in figure 3.2. In this example, the dataset consists of
A’s and B’s with different characteristics. A decision tree is then used to classify the data set
into classes. In this case, it classifies the dataset based on color and whether the letter is
underlined or not. When splitting at a node in a normal decision tree, the feature which splits
the set into the biggest difference in the next right and left node from the initial set is always
chosen (Hastie, Tibshirani & Friedman, 2008; Yiu, 2019). The first split is happening at the
top, at the “Is black?” question. Here, the dataset is split into two separate classes where one
consists of black letters and the other one red letters. All the letters to the right are all the
same, so the classification is done on the right side in the tree. To the left, it is possible to see
that there are A’s which are underlined and not, so these are split by the question “Is
underlined?”. This resulted in two new classes, one with A’s underlined and one where the
A’s does not have an underline. This is a simple overview of decision trees, and in practice
they can be much more complicated.

![Figure 3.2: Overview of a simple decision tree. (Based on Yiu, 2019)](image)

The random forest model uses the individual trees within the forest by letting each tree make
a class prediction, then putting together all the answers from each individual tree to make the
random forest model prediction. See figure 3.3 below:
In figure 3.3, there are 4 trees within the forest where each tree predicts A or B. All the predictions in the trees in the forest are then counted, and the prediction which has the highest count becomes the random forest prediction. Which was A in the figure above.

To prevent that each tree’s behavior is similar to each other, two techniques are used. The first one is called bootstrap aggregation or bagging, and ensures that each tree in the random forest is trained on different datasets (Hastie, Tibshirani & Friedman, 2008; Yiu, 2019). This is done in the following way: Each tree randomly choose a sample from the dataset which is used by the random forest. Each sample is the same size as the original dataset, but with the difference that duplicates can occur. See figure 3.4 below:
The second technique is called feature randomness. Recall that a normal decision tree always splits a node which creates the biggest separation between the newly created left and right node. With feature randomness, the trees in a random forest selects a random chunk of features and then selects the feature from the random chunk which creates the biggest separation between the newly created left and right node (Hastie, Tibshirani & Friedman, 2008; Yiu, 2019). In figure 3.5 below, an example with 4 features is illustrated. In the trees to the right, which belongs to a random forest, two features are randomly selected, and the feature which creates the biggest separation between the left and right node is chosen. In the figure below, the feature that is chosen is bold and underlined.
The amount of trees and how many features that can randomly be selected by each tree can be altered to try to improve the results. Oshiro, Perez & Baranauskas (2012) argues that between 64 and 128 trees in a forest is enough in most cases, and that the difference between 128-4096 trees is so small that it is not worth the computational power. In the experiment in this paper, the number of random features chosen were 2 and the number of trees were 500. The number of random features were changed to 3 and 4 to see if any improvements in the results could be achieved, no improvements could be seen and therefore the number of two features were chosen.

3.3.1.2 Support vector machine

A support vector machine solves classification problems simply by finding a line, called a hyperplane, that separates input data into different data classes (Pupale, 2018; Hastie, Tibshirani & Friedman, 2008). See figure 3.6 below.

![Figure 3.6: A visualisation how a SVM separates a dataset with a line. (Based on Starmer, 2019)](image)

The problem is to a hyperplane that separates the data set in an optimal way. In the example above, where the classes (orange and purple dots) are clearly separated, this is done by finding the points within each separate class which are closest to the hyperplane. When those points are found, the distance between these points and the hyperplane is calculated, and the goal is to maximize this distance (Pupale, 2018). This hyperplane is called a maximum margin classifier (Hastie, Tibshirani & Friedman, 2008). If new data was added, that data could now be classified as an orange or purple dot. See figure 3.7 below.

![Figure 3.7: A visualisation of how the hyperplane is decided. (Based on Starmer, 2019)](image)

Usually, the data is not as separated as the example above. Looking at the example below in figure 3.8, if new data was added (the black dot) it would be classified as belonging to the orange class, even though it is far more close to the bundle of purple dots at the right side.
To solve this problem, SVM uses two different techniques. The first one is allowing misclassifications and is illustrated in figure 3.9 below.

Misclassification is allowing for data points to be ignored when the hyperplane is drawn. So in the example in figure 3.9 above, the new black dot will be classified as purple because the transparent orange dot between the black and the purple won’t be used when calculating the hyperplane. The distance between the observations and the hyperplane is called the soft margin or support vector classifier and can be seen in figure 3.9. SVM uses cross validation to decide how many misclassifications and observations that are allowed to be within the soft margin. (Hastie, Tibshirani & Friedman, 2008)

The second technique that is used by SVM is when a problem is not linearly separable. However, when a dataset cannot be linearly separable, SVM can transform the data into a higher dimension, to be able to separate it in that space (Pupale, 2018; Hastie, Tibshirani & Friedman, 2008). See example in figure 3.10 below.

To solve this, each data can be transformed into a 2 dimensional space by taking each point in figure 3.10 and calculate the square value. The results is visualised in figure 3.11.
So a SVM tries to find a line which classifies the data by allowing misclassifications with support vector classifiers and if the data is inseparable then recalculates the data to represent it in a higher dimension.

3.3.2 Multiple logistic regression model

The multiple logistic regression (MLR) model was chosen because it is a traditional, popular model to use for classification problems which also makes it possible to use several independent variables and one dependent variable, a so called many-to-one relationship (Knofczynski, 2008; Yilmaz & Kaynar, 2011; Allison, 1999; Menard, 2001). The MLR uses the independent variables to try to predict the dependent variable, hence the name of the dependent variable. The goal is to model the connection between the independent variables and the dependent variable (Menard, 2001). This is done by presenting a dataset with both independent variables and their corresponding dependent variable so that patterns between variables can be established with training. When that is done, the MLR can be used with only independent variables to try to predict their corresponding dependent variable. The MLR equation that is used is as follows:

\[
\log(\text{odds}) = \logit(p) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + \epsilon
\]

Where

- \( p \) is the probability that a person did not quit their job
- \( x_i \) are the independent variables
- \( \beta_0 \) is the expected value on the prediction when all independent variable weights are 0
- \( \beta_{1-p} \) are the weights for each independent variable, called regression coefficients
The equation above is used to calculate the log odds ratio. The log odds ratio is another expression for probability and odds ratio (Peng, Lee & Ingersoll, 2002). In this case, it is the probability of employees quitting their job. Each independent variable weight decides how much that specific independent variable affects the log odds ratio, when the independent variable is changing (Allison, 1999; Menard, 2001). That means that a big weight for an independent variable is more important for the determination of the prediction. Each weight is the log of the odds ratio for that specific independent variable, and the odds ratio for each independent variable can be calculated by $exp(\beta_{1-p})$. The odds ratio for each independent variable is a way to show how the prediction changes for each unit change in one independent variable. The weights are obtained by giving the MLR model a dataset (similarly to presenting a training data set to the ML models) to train on, to find the patterns which was mentioned earlier.

3.3.2.1 Usage of MLR in this study

The programming language R (version R 3.6.3 with GUI version 1.70) and the package “caret” was used to train, generate the model equation and to test the accuracy of the MLR model. Five different tests were made with the MLR model. In each test, the model was trained with a different dataset. The datasets that were used was the same as in the machine learning models testing and can be found in section 3.3. The variables presented as “input variables” was used as independent variables and the “output variable” was used as dependent variable.

Each dataset were split into two partitions. One with 20% and the other with 80% of the dataset. The model was then trained with the partition containing 80% of the dataset. When the training was made, the coefficients which was used in the model was extracted with a caret function in R. This was done to be able to show the specific equation that the models calculated and used. The model was then used on the remaining 20% of the data to test the accuracy.

3.4 Quality of research

In order to achieve high quality and trustworthiness it was necessary to assess the validity and reliability of this paper (Gibbert, 2008; Yin, 1994). Validity is split into three parts; internal validity, external validity and construct validity. Internal validity refers to the connection between variables and results. High internal validity is achieved when no variables that was not thought of have an impact on the result (Gibbert, 2008; Yin, 1994). The methodological choices made was carefully considered, and presented in the method section and accounts for why no external, not thought of reasons, had an impact on the result. And that contributed to high internal validity. The external validity refers to how the results can be generalized to other settings than just the one that the research was conducted in (Gibbert, 2008; Yin, 1994;
McGrath & Brinberg, 1983). To achieve this, the interviews were conducted with employees
had different roles and worked for several different organizations in different branches to
ensure that the result was not only relevant for the case company. The construct validity
refers to how the data is collected and how it is treated. To achieve construct validity,
multiple sources of information and data collection needs to be used and the reader has to be
able to understand how the researcher went from research questions to a conclusion (Gibbert,
2008; Yin, 1994). This was achieved throughout the whole process of the paper and was
presented in the method section.

Reliability refers to transparency and how replicable the research is (Gibbert, 2008; Denzin
& Lincoln, 1994). According to Gibbert (2008) all the procedures that has been made during a
research has to be documented to achieve transparency, which is exactly what has been done
in the method section within this paper. The fact that the procedures were well documented,
also ensures that the study is replicable.

Also, to achieve high quality, four guidelines from Vetenskapsrådet (2019) were followed.
These were confidentiality requirement, consent requirement, good use requirement and
information requirement. The confidentiality requirement means that all the data that was
collected was stored in such a way that unauthorized people could not get access to it, and to
that all the persons that were involved in the study (the interviewees in this case) was ensured
to be anonymous. The consent requirement means that all of the participants that somehow
was a part of this study, had the choice to decide whether they wanted to participate or not.
The good use requirement means that all of the data that was gathered, were only used for the
purposes that was communicated to the participants which the information was gathered
from. The information requirement means that all of the participants were informed about the
purpose and what the data gathered would be used for.
4 Empirics

In the empirics section, all the relevant gathered findings from the machine learnings experiments as well as the findings from all of the interviews are presented. The section is split up in three parts, the first one where all the results from the machine learning experiments conducted in R are presented. The second part is where the multiple logistic regression model results are presented. The third part consists of the findings from the interviews. The presented results in the third part are from interviews with HR specialists, line managers and employees whom are working beneath line managers.

4.1 Random forest and SVM results

The results from the two different machine learning algorithms, random forest and support vector machine, are presented in tables down below. The random forest algorithm will from here be named as Rf and the support vector machine will be named as svm. There are five different results for each algorithm because of the five test cases, where the partition between people who quit and people who stay differs and the amount of data points differ. The circumstances for each test will be presented to each test result table. Specific information such as which data points that were used can be seen in section 3.2.2.1. In the tables in this section, the column Algorithm describes which algorithm was used, column Accuracy % is the accuracy which the results showed that both models achieved. Accuracy in this case means in how many test cases that the algorithm is predicting the right answer, based on a validation set of the data.

In the following six subsections, the results from the respective test 1-5 will be presented, as well as a combined summary of these tests.

4.1.1 Model comparison

In table 4.1 we can see a representation of the total amount of people and the partition between people who have quit their job and people who has not quit their job in all of the five tests that were made with the random forest and support vector machine models. So for example, for the first comparison, 16640 people did not quit their job and 484 did. Recall from section 3.3.1 that 80% of the dataset used in each comparison was used for training, and 20% was used for accuracy testing, which we can see the results of in table 4.2. For comparison 1-3, six input data points and one output data point, whether a person quit their job or not, were used. And for the last two comparisons (test 4 and 5), a time perspective was added so that the number of input data points were 12. Apart from the six original data points, changes in clarity, value, efficiency, workload, community and enthusiasm between two different pulse surveys were added.
In table 4.2 down below, the accuracy of both the random forest and for each tests are presented. We can see that the accuracy in each test are very close to the percentage of how the input data sets were partitioned with people who quit their job and people who did not quit their job. We could also see that the results were the same, even though the input data sets changed. The results will be further discussed and analyzed in section 5.

<table>
<thead>
<tr>
<th>Test</th>
<th>Have quit the job</th>
<th>Amount of people</th>
<th>Percentage %</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>no</td>
<td>16640</td>
<td>97.17</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>484</td>
<td>2.83</td>
</tr>
<tr>
<td>2</td>
<td>no</td>
<td>10000</td>
<td>88.25</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>1331</td>
<td>11.75</td>
</tr>
<tr>
<td>3</td>
<td>no</td>
<td>3873</td>
<td>74.42</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>1331</td>
<td>25.58</td>
</tr>
<tr>
<td>4</td>
<td>no</td>
<td>16640</td>
<td>97.17</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>484</td>
<td>2.83</td>
</tr>
<tr>
<td>5</td>
<td>no</td>
<td>4620</td>
<td>90.48</td>
</tr>
<tr>
<td></td>
<td>yes</td>
<td>484</td>
<td>9.52</td>
</tr>
</tbody>
</table>

Table 4.1: The partitioning between people that quit and did not quit their job in test 1-5

<table>
<thead>
<tr>
<th>Test</th>
<th>Algorithm</th>
<th>Accuracy %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test 1</td>
<td>Rf</td>
<td>97.17</td>
</tr>
<tr>
<td></td>
<td>svm</td>
<td>97.17</td>
</tr>
<tr>
<td>Test 2</td>
<td>Rf</td>
<td>88.30</td>
</tr>
<tr>
<td></td>
<td>svm</td>
<td>88.30</td>
</tr>
<tr>
<td>Test 3</td>
<td>Rf</td>
<td>74.77</td>
</tr>
<tr>
<td></td>
<td>svm</td>
<td>74.73</td>
</tr>
<tr>
<td>Test 4</td>
<td>Rf</td>
<td>97.17</td>
</tr>
<tr>
<td></td>
<td>svm</td>
<td>97.17</td>
</tr>
<tr>
<td>Test 5</td>
<td>Rf</td>
<td>90.48</td>
</tr>
<tr>
<td></td>
<td>svm</td>
<td>90.48</td>
</tr>
</tbody>
</table>

Table 4.2: A combined summary of the accuracy of both algorithms for each test.
4.2 Multiple linear regression model results

The same five datasets that were used for testing the SVM model and the RF model were used for the multiple logistic regression model. How the data sets were partitioned with people who quit their job and did not quit their job can be seen in table 4.2. The independent variables for test 6-8 were clarity, value, efficiency, workload, community and enthusiasm. For test 9-10, all of the above variables were used and the change in these variables between two surveys were added. The dependent variable were the same for all tests, whether a person quit their job or not.

4.2.1 Test 6

In the first test for the MLR model, dataset 1 was used. This dataset contained 17124 rows, which is 17124 sets of independent variables and one corresponding dependent variable. The coefficients for the model and the model equation can be found below. The accuracy was measured using the remaining 20% of the dataset that was not used for training. The test showed an accuracy of 97.17%.

<table>
<thead>
<tr>
<th>Variable</th>
<th>( \beta_p ) (weight)</th>
<th>Odds ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_0 ) (constant)</td>
<td>-2.779</td>
<td></td>
</tr>
<tr>
<td>Clarity</td>
<td>-0.255</td>
<td>0.775</td>
</tr>
<tr>
<td>Value</td>
<td>0.118</td>
<td>1.125</td>
</tr>
<tr>
<td>Efficiency</td>
<td>0.047</td>
<td>1.048</td>
</tr>
<tr>
<td>Workload</td>
<td>0.098</td>
<td>1.103</td>
</tr>
<tr>
<td>Community</td>
<td>-0.088</td>
<td>0.916</td>
</tr>
<tr>
<td>Enthusiasm</td>
<td>-0.139</td>
<td>0.870</td>
</tr>
</tbody>
</table>

*Table 4.3: Regression coefficients for test 6 and their individual weight*

From the extracted regression coefficients, the regression equation that was used by the model was:

\[
y \approx -2.779 - 0.255(\text{Clarity}) + 0.118(\text{Value}) + 0.047(\text{Efficiency}) + 0.098(\text{Workload}) \\
- 0.088(\text{Community}) - 0.139(\text{Enthusiasm})
\]

4.2.2 Test 7

In this test, the total amount of people were 11331. The amount of people and the partitioning between people who quit their jobs and did not quit their jobs was the only change in input data from test 7. The odds ratio and the weights for each independent variable can be found in table 4.4. The equation corresponding the model can be found below. The accuracy test was
made on the remaining 20% of the dataset that was not used for training. The results showed an accuracy of 88.25%

<table>
<thead>
<tr>
<th>Variable</th>
<th>( \beta_p ) (weight)</th>
<th>Odds ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_0 ) (constant)</td>
<td>-1.600</td>
<td>0.848</td>
</tr>
<tr>
<td>Clarity</td>
<td>-0.165</td>
<td>1.019</td>
</tr>
<tr>
<td>Value</td>
<td>0.019</td>
<td>1.129</td>
</tr>
<tr>
<td>Efficiency</td>
<td>0.121</td>
<td>0.975</td>
</tr>
<tr>
<td>Workload</td>
<td>-0.025</td>
<td>1.134</td>
</tr>
<tr>
<td>Community</td>
<td>0.126</td>
<td>0.822</td>
</tr>
<tr>
<td>Enthusiasm</td>
<td>-0.196</td>
<td>0.854</td>
</tr>
</tbody>
</table>

*Table 4.4: Regression coefficients for test 7 and their individual weight*

From the extracted regression coefficients, the regression equation that was used by the model was:

\[
y \approx -1.600 - 0.165(\text{Clarity}) + 0.019(\text{Value}) + 0.121(\text{Efficiency}) - 0.025(\text{Workload}) + 0.126(\text{Community}) - 0.196(\text{Enthusiasm})
\]

4.2.3 Test 8

The input data was changed again from test 6 and 7. In this test, the total amount of people were 5204. Weights and odds ratio can be found in the table below. The accuracy test showed an accuracy of 74.44%.

<table>
<thead>
<tr>
<th>Variable</th>
<th>( \beta_p ) (weight)</th>
<th>Odds ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_0 ) (constant)</td>
<td>-0.630</td>
<td>0.855</td>
</tr>
<tr>
<td>Clarity</td>
<td>-0.156</td>
<td>1.011</td>
</tr>
<tr>
<td>Value</td>
<td>0.012</td>
<td>1.121</td>
</tr>
<tr>
<td>Efficiency</td>
<td>0.115</td>
<td>0.946</td>
</tr>
<tr>
<td>Workload</td>
<td>-0.055</td>
<td>1.117</td>
</tr>
<tr>
<td>Community</td>
<td>0.111</td>
<td>0.854</td>
</tr>
<tr>
<td>Enthusiasm</td>
<td>-0.157</td>
<td>0.854</td>
</tr>
</tbody>
</table>

*Table 4.5: Regression coefficients for test 3 and their individual weight*

From the extracted regression coefficients, the regression equation that was used by the model was:

\[
y \approx -0.630 - 0.156(\text{Clarity}) + 0.012(\text{Value}) + 0.115(\text{Efficiency}) - 0.055(\text{Workload}) + 0.111(\text{Community}) - 0.157(\text{Enthusiasm})
\]
4.2.4 Test 9

In this test, the input data was changed in several ways. In this test, the total amount of people were 17124 and 6 additional independent variables were added compared to test 6-8. These are the change in clarity, value, efficiency, workload, community and enthusiasm between two surveys. They are called $\Delta$ and the variable name in the table down below. These were added to see if there were any interesting change in the results when the aspect of time was added. The weights and odds ratios for each variable can be found in the table below. The equation that was extracted from the model can be found beneath the table. The accuracy for this test was measured to an accuracy of 97.15%.

<table>
<thead>
<tr>
<th>Variable</th>
<th>$\beta_p$ (weight)</th>
<th>Odds ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$ (constant)</td>
<td>-4.816</td>
<td>0.776</td>
</tr>
<tr>
<td>Clarity</td>
<td>0.254</td>
<td>1.144</td>
</tr>
<tr>
<td>Value</td>
<td>0.135</td>
<td>1.033</td>
</tr>
<tr>
<td>Efficiency</td>
<td>0.033</td>
<td>1.137</td>
</tr>
<tr>
<td>Workload</td>
<td>0.128</td>
<td>0.944</td>
</tr>
<tr>
<td>Community</td>
<td>-0.057</td>
<td>0.862</td>
</tr>
<tr>
<td>Enthusiasm</td>
<td>-0.149</td>
<td>0.974</td>
</tr>
<tr>
<td>$\Delta$ Clarity</td>
<td>-0.026</td>
<td>1.084</td>
</tr>
<tr>
<td>$\Delta$ Value</td>
<td>0.080</td>
<td>1.089</td>
</tr>
<tr>
<td>$\Delta$ Efficiency</td>
<td>0.085</td>
<td>1.071</td>
</tr>
<tr>
<td>$\Delta$ Workload</td>
<td>0.068</td>
<td>1.069</td>
</tr>
<tr>
<td>$\Delta$ Community</td>
<td>0.066</td>
<td>0.988</td>
</tr>
<tr>
<td>$\Delta$ Enthusiasm</td>
<td>-0.012</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.6: Regression coefficients for test 9 and their individual weight

From the extracted regression coefficients, the regression equation that was used by the model was:

$$y \approx -4.816 + 0.254(\text{Clarity}) + 0.135(\text{Value}) + 0.033(\text{Efficiency}) + 0.128(\text{Workload})$$

$$- 0.057(\text{Community}) - 0.149(\text{Enthusiasm}) - 0.026(\Delta \text{Clarity}) + 0.080(\Delta \text{Value})$$

$$+ 0.085(\Delta \text{Efficiency}) + 0.068(\Delta \text{Workload}) + 0.066(\Delta \text{Community}) - 0.012(\Delta \text{Enthusiasm})$$

4.2.5 Test 10

This test was made with similar conditions as test 9. The only change was the partitioning in the outcome and the amount of people. In this test, the total amount of people were 5106. Weights and odds ratios for each variable are presented down below, as well as the corresponding equation to the model. The accuracy test showed an accuracy of 90.50%.
Table 4.7: Regression coefficients for test 10 and their individual weight

<table>
<thead>
<tr>
<th>Variable</th>
<th>$\beta_p$ (weight)</th>
<th>Odds ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$ (constant)</td>
<td>-3.594</td>
<td>0.751</td>
</tr>
<tr>
<td>Clarity</td>
<td>-0.287</td>
<td>1.197</td>
</tr>
<tr>
<td>Value</td>
<td>0.180</td>
<td></td>
</tr>
<tr>
<td>Efficiency</td>
<td>-0.031</td>
<td>0.969</td>
</tr>
<tr>
<td>Workload</td>
<td>0.147</td>
<td>1.158</td>
</tr>
<tr>
<td>Community</td>
<td>-0.074</td>
<td>0.929</td>
</tr>
<tr>
<td>Enthusiasm</td>
<td>-0.138</td>
<td>0.871</td>
</tr>
<tr>
<td>$\Delta$ Clarity</td>
<td>0.092</td>
<td>1.096</td>
</tr>
<tr>
<td>$\Delta$ Value</td>
<td>0.089</td>
<td>1.093</td>
</tr>
<tr>
<td>$\Delta$ Efficiency</td>
<td>0.075</td>
<td>1.078</td>
</tr>
<tr>
<td>$\Delta$ Workload</td>
<td>0.061</td>
<td>1.063</td>
</tr>
<tr>
<td>$\Delta$ Community</td>
<td>-0.016</td>
<td>0.985</td>
</tr>
</tbody>
</table>

From the extracted regression coefficients, the regression equation that was used by the model was:

$$y \approx 3.594 - 0.287(Clarity) + 0.180(Value) - 0.031(Efficiency) + 0.147(Workload)$$

$$- 0.074(Community) - 0.138(Enthusiasm) - 0(\Delta Clarity) + 0.092(\Delta Value)$$

$$+ 0.089(\Delta Efficiency) + 0.075(\Delta Workload) + 0.061(\Delta Community) - 0.016(\Delta Enthusiasm)$$

4.3 Interviews

The interviews were conducted mainly to answer and contribute to RQ 2 and 3, namely: (2) *How does understanding a ML model affect people’s will to use them?* and (3) *What are practitioners perceptions and beliefs about how organizational performance could be affected by being able to predict employee attrition on an individual level?*. Presented results are the most common and the most interesting answers regarding the research questions. Only empirics gathered through interviews are presented in this section.

4.3.1 The importance of trust and understanding a ML model

All of the interviewees said that the trust in a ML model which is supposed to make assumptions about life changing decisions played a huge part for their will to both use, and to be a part of a model like that. As interviewee F said: “The consequences could be too big so for decisions like that, I need to be able to trust the model to be fair in the judgement it makes of me, otherwise I would not want a model to judge me”. Several of the interviewees mention themselves and their will to use ML models for life changing decisions. When asked what they think about other people’s thoughts in using such models, they believe that most of the people think that a model needs to be trusted to be used. That is confirmed by the fact that
each interviewee said that trusting a model is essential for its use. This is also strengthened by the literature in section 2.3.1 where Ribeiro, Singh & Guestrin (2016) said that people would not be open to use such models if they did not trust them.

However, trusting and understanding a model was shown to not be the same. Only 2 out of 10 interviewees thought that interpretability was important where interviewee L said: “If I’m not able to understand how a model works, how am I supposed to trust it?” On the contrary, 8 other interviewees said that they personally would not need to be able to understand exactly how a model works, but that someone that they trust need to be standing behind a model for them to trust the model itself. That someone could be a person or an organisation as a whole. Interviewee I said: “Personally it could be hard for me to understand machine learning models, and I’m not that interested in machine learning or AI for that matter as well. But for me, it could be a friend that is telling me that it is fine or that if an organization which i trust uses a model like that, I would be okay with it as well.”. Interviewee H fills in: “… as i understand it, like, companies like instagram use this kind of technology for advertising, but yeah, I don't really care … but for life changing decisions, hmm, yeah as long I trust the organisation I could see myself trusting a model like that.”

4.3.2 Organizational impact
All of the interviewees agree that a machine learning method that can be used to predict employee attrition on an individual level could have positive effects on an organisation in several ways. 9 out of 10 mentioned that the information that a specific employee would quit in a certain amount of time could be used proactively in several ways. Firstly, it could be used to start a recruitment process in far earlier stages, looking for replacement for that particular role. Interviewee D said “It would be amazing for HR departments. We could in much earlier stages develop trainee programs to find and educate the right person, but not only that, just to be on the market for new employees at an earlier stage would be beneficial.” and D continues: “…it could minimize the negative impact on the transition between two employees”. Several other interviewees mention the same thing as well. Interviewee F said: “Organizations could start looking for new recruits in an earlier stage. The earlier we can start, the bigger chance to find the right person … It would be smoother workflow when an employee quits and a new person replaces, minimizing the risk for productivity to go down or at least not as much as it would if we worked reactively”.

6 interviewees also talked about knowing when people are quitting, and when they actually are quitting, should raise the question “Why are people leaving our organization?”. Interviewee C said: “If we can see with a model that people are leaving our organization, we have to ask the question why. With that information, we could start looking into factors that contribute to that and take action based on what the findings are.”. And that information could minimize the chances of people leaving in the future. All six agree that in the long
term, this could have a positive effect. However, 3 of them think that the information gained by such an investigation, to why a person want to leave, should not be used to try to make the person stay in the particular case, but rather looking at a more general level what the organization could do to prevent people from leaving in the future. Interviewee F said: "If a person has decided to quit, I believe that for that person to stay, big changes need to occur, and that might affect other employees negatively which could have big consequences. But if there is a pattern that people are leaving, then it is perfect to look into the underlying reasons".

There were 6 interviewees who talked about how people quitting their jobs could affect the spirit and culture in an organization. C said: "The culture and relations in organisations is important. The people are the ones who are delivering the results and there is a risk that the culture can break if a person quits. And that could have a negative impact on the productivity". They also agreed that this was a bigger problem for smaller organizations or smaller units within an organisation. Interviewee D said: "Our business is based on relations, and if one person leaves, there is a high chance that several will leave. This could be devastating for our organisation.". And F continues: "... well if a unit of an organization quits, then it could be a problem. But if individuals quit, long term it should not be a problem if the organization is not very small.". Within the same area, 5 interviewees mention that relations between employees are important for the productivity and that if people are leaving, these relations can break. And one interviewee also said that relations can break, but hiring new people could also build new good or even better relations than before. All of the interviewees who were working within HR also mentioned that a problem with people leaving their organizations was the loss of knowledge. Interviewee D said "It could definitely have a negative effect on our ability to produce what we do, specially if there are several people leaving. Their knowledge is seldom replaced simply by hiring a new person to that position and often it takes time for a new person to get up to speed."

Out of 8 persons who thought that using a model to predict employee attrition was fine, 6 said that decisions should not solely be based on the results of such a model, and the other two did not mention it at all. The model should rather work as information gathering, and as a complement to dialogues. They mention that the most important part for the usage of such a model, and for decisions like quitting a job, should be to create awareness and start dialogues with the employees. The results of a model can work as an indicator that could propose for managers and HR departments to start investigating the results of a model by simply starting a dialogue with the employee. Interviewee C said: "I never think that the model should be used for decision making, but rather work as an indicator that okay something is wrong here and we have to look into it further before we can actually take action."
4.3.3 Ethical dilemmas

There were not only positive organizational effects that were brought up during the interviews and the majority of the interviewees also mentioned situations where it could be problematic to use a model that predicts employee attrition. Interviewees A, C, B, F and I said that if information about a person quitting a job is acquired, even before the employee knows that he will quit, then this could lead to that their managers could change the behaviour in a way which affect the person negatively. For example interviewee A said: “… because if the manager knows that a person will quit, it is likely that his behaviour would change for the worse towards the employee. I mean, imagine if they were to send people on education, why would they send a person that a computer would say that the person would quit? Even if the employee had not said that himself … and it could be as simple as stop communicating with the employee”. So the prioritization towards employees could shift towards those who the organisation knows will stay.

There is also the matter the sensitivity of the information and several ethical aspects were mentioned during the interviews. Just as the case was in the examples from the interviews above, if organizations, managers and employees start to treat a person who is quitting, according to a model, unfair and also could be ethically wrong. Interviewee A mention: “What if the model is telling the organization or the person who is in charge that a person will quit. And then they start to treat this person in an unfair way, unconsciously or consciously, and the model is wrong. The employee did not think of quitting, but because of how the other peoples treatment have changed, the employee actually quits.”. In a case like this, the organization could wrongly force someone or at least change peoples mind, almost in a manipulative way. Also, 2 interviewees mention that an employer could use information gained from a model such as this to their own advantage in negotiating with their employees. For example C said “You are negotiating with your boss and if a model that can tell the boss when a person quits or not is used, then he gets the upper hand and that is unfair.”

Half of the interviewees said that the only people who should have this information is the HR department, and that they have to be completely unbiased between the employer and employee. And then it is their task to start a dialogue with the employee to see how the situation actually is. The reason for that is the simple fact that otherwise, the employer could use it only for their own good, and that could put the employee in a bad spot. The other half of the interviewees thought that the information should be transparent between all involved parties, both HR, managers and the employee.

Transparency is important and 7 of the interviewees mentions transparency as a key factor to usage of a model that can predict employee attrition. As long as all parts are transparent to each other, and are open with how the model works, how it is used and for what purpose, then it would be okay to use it. However, as mentioned above, there is a difference between
having transparency about that a model is used, and transparency about the results. But if an organization uses a tool like this to gain more knowledge about their employees to gain an advantage, then all seven who mentions transparency think that it should be forbidden and that they would not comply with a such a model.

Another issue that was brought up during the interviews was side effects with using algorithms for decisions. There were different opinions. Interviewee B said “What if for example we are using an algorithm for hiring people, and it is biased towards one particular group of people? We could miss out on very interesting candidates”. At the same time, another interviewee mentions that exact same situation and interviewee I a similar one, but they think that algorithms would be better than how it is today. Interviewee D said “You have to be a really really sharp recruiter to not be biased in a recruitment process. Your own values does play a big part, so I think that computers can be better in that sense even if they are not perfect.”.

Even though all interviewees mention difficult ethical aspects that needs to be considered, 8 out of 10 would be okay with their organizations using models which could predict employee attrition. They also believe that a big part of the population would comply and that we are living in an era where technology is accepted as a tool rather than something that is used against us. HR specialist D said that “This would be very welcomed. It would give us advantages in handling employee turnover, and I am certain that almost everyone would welcome a model like this.”. On the contrary, one interviewee was strongly against it and said: “On a personal level? That would violate people's privacy. That is not accepted so no, I can't imagine where this would be okay to use,“
5 Analysis and discussion

This section contains the analysis of the empirics that was found, based on theory, in order to be able to answer the research questions that was introduced in the beginning of the paper. Only the most important and interesting findings are discussed and analyzed. First, an analysis of the models and how well they they could predict employee attrition is presented and secondly, the empirics of the interviews are discussed.

5.1 Random forest and Support vector machine

Recall that both the random forest and the support vector machine were used as supervised algorithms which were presented in section 2.1.1, where the goal is to be able to predict output based on learning from a training dataset where training data consisted of both input and output so that the algorithm could learn the patterns and use those patterns to predict new outcome.

The accuracy of both the random forest model and support vector machine model were almost identical, and just reading the numbers indicates that both models were very accurate. In the first and fourth test, the models showed an accuracy of 97%. However, analyzing the results further, we could see that the accuracy was the same as the partitioning on people who quit their job and people who did not quit their job. This raises a question whether both models predicted that all people would not quit their job. Looking at the results from all 5 tests for both models, we could see that the pattern looks the same. All results are almost identical to the partitioning with the input dataset. Through a sample it could be seen that each test did not predict that all of the employees would not quit their jobs.

5.1.1 Impact of partitioning in the used dataset

One reason for this could be explained by the input dataset partitioning itself. If the input dataset contains 97% of people who did not quit their job and 3% who did quit their job, then could a model possibly predict that 100% did not quit their job and achieve an accuracy of 100%. This is a well known problem within the field of machine learning and is called overfitting (Hawkins, 2004; Hastie et al. 2008; Srivastava, 2014). And this happens when the model learns to predict the training data too well. The risk of overfitting could be lowered in several ways in which some of them are with using cross-validation, train with more data or regularization (Hawkins, 2004; Srivastava, 2014). To try to prevent overfitting, the partitioning was changed and tested again, however the results still showed the same pattern. With the datasets used for training might have been overfitted towards employees who did not quit their job. On the contrary, it could have been needed to overfit the data even more towards that employees quit their job. To some extent that was done, but not necessarily enough. However, overfitting too much towards employees who quit their job would change
the fact that the tests was tried with real data. Altering the datasets would require to remove people who did not quit their job. This could however potentially be a good path to go, since the purpose was to predict when people quit, and not the other way around. Indirectly it is the same but there are two distinct goals which would be two different problems for the algorithms to solve.

5.1.2 Input variable impact

Another reason could be the chosen input variables and how much they actually affected the predictions. The chosen input variables were clarity, value, efficiency, workload, community and enthusiasm, and their change over two different pulse survey occasions. These were chosen because they were deemed to affect the factor job satisfaction which plays a big role in why people choose to stay or leave their organizations according to the theory which was presented in section 2.1.3 (Mitchell et al. 2003; Hausknecht, Rodda & Howard, 2009). These variables could very well be the right variables to look at, however the way that they are measured might imply implications for the model when it is trained. The four different values that each variable can be, 0-3, could possibly prevent the model from changing the prediction in a significant way, when a variable changes.

Another matter that has to be discussed regarding the input variables and the dataset itself is the fact that the first three datasets were only based on one pulse survey. In most cases when an employee leaves an organization voluntary, it’s not a decision who just shows up over night but rather this grows over time. So for these tests to actually show something interesting in regards of predicting employee attrition seems far away and almost impossible based on that fact.

Therefore, the time aspect was added for test number 4 and 5. This was done by adding the change in clarity, value, efficiency, workload, community and enthusiasm between two separate pulse survey occasions. The time between those two pulse surveys were two months. The results in the fourth and the fifth test yielded the same results as test one, two and three, namely that the accuracy of each model was the same as the partitioning within the training dataset. Two months between the pulse surveys that were compared might have been too short to yield interesting results. The longer the timespan the bigger chance of getting a difference that matters. Also, only adding two pulse surveys might have been too few. Adding more pulse surveys would provide the models with more accurate data over time, and since time is a factor of when employees grow a decision to leave an organisation, it is likely that the time factor needs to be considered in a more significant way than in these tests.

Another interesting thought to include here is that the usual mutual time of notice for an employee when quitting a job in Sweden is between one to three months (Unionen, 2020). This fact might have implications on all of the above tests. Since the datasets used are based
on pulse surveys that was made one month to three months before an employee leaves the organisation, it might be the wrong point in time to look at pulse survey results. The actual pulse survey that would be interesting to look at might occur several months before they actually quit their job, even a few months before they make the decision to go to their employer and tell them that they are quitting. If the employee answers the surveys after the decision to quit was made, it is possible that the measures in the survey might go up and be better and therefore also affect the prediction. Neuman (2007) argues that the health, happiness and satisfaction might improve when people voluntarily quit their job. So the pulse survey that would be the most interesting to look at is the one where the employee make the final decision that they are quitting and not the 1-3 pulse surveys closest to when they actually quit. In all the tests made in this paper, that is not the case. This is also one reason that the time aspect might need to be considered on a deeper level to accomplish good results with these models.

5.2 Multiple logistic regression model

The circumstances for tests with the MLR model was the same as the tests with the ML models. In terms of accuracy, we could see that the results were almost identical as the tests with the ML models. Each test received an accuracy that was similar to the percentage of partitioning between people who quit and did not quit their jobs in the dataset that was used for training. The discussions made in section 5.1.1 “Impact of partitioning in the used dataset” and 5.1.2 “Input variable impact” also holds for the tests that were made with the MLR model.

Another factor for that could have played a part in the results could be that the independent variables are too highly correlated to each other. A phenomenon called multicollinearity. Multicollinearity is occurring when the independent variables have to high correlation between each other (Alin, 2010; Graham, 2003). The consequences could be that changes in one independent variable could infer changes in other independent variables so that the predicted value does not change enough for the result in the prediction to change or that the prediction changes in an unrealistic way (Alin, 2010). It could also be hard to acknowledge how changes in one independent variable affects the predicted outcome. So for example, suppose we have variables a and b and the prediction c, a and b determines c. Imagine the equation \( a + b = c \) and let \( a = 1, b = 1 \) then \( c = 2 \). Now imagine \( a = 2 \) and that \( b \) is correlated to \( a \) in a way so that when \( a = 2 \) then \( b = 0 \). This would give the same answer \( c = 2 \) as in the first case. This takes away that the regression coefficient explains how the predicted value is changing when the corresponding independent variable is changing. The result is instead that the prediction might change, but we cannot be certain that the prediction is changed due to the change in one independent variable but instead in several.
We could see in table 4.3, 4.4, 4.5, 4.6 and 4.7 that the weights to the independent variables varies between positive and negative. This means that the positive weights moves the predicted value in one direction and the negative in the opposite direction. Looking specifically at the variable called community throughout all five tests for the MLR method, we can see that they shift between positive and negative. Recall from section 3.2.2.1 that community refers to how well an employee feel that they belong at work. That knowledge and looking at the weights for that variable between the tests, raises the question of how could it be positive in one test and negative in another? According to Mitchell et al. (2003) and Hausknecht, Rodda & Howard (2009), community is a parameter that plays a big role whether people quit or not from their jobs. Therefore, the weight on the community variable should have the same sign throughout all of the tests. This is an indication that the tests did not result in accurate answers and that the community variable did not affect the predicted value in a significant way.

5.3 Interviews

5.3.1 Importance of trust and interpretability

In section 2.3.1 it was brought up that according to Ribeiro, Singh & Guestrin (2016) people would not be open to use ML models that were used for life changing decisions if they could not fully trust them. The empirics in section 4.3.1 strengthens that theory. We could see that all of the 10 interviewees that participated in this research, agreed that trusting a model would be essential for their will to somehow use such a model. Just as Ribeiro, Singh & Guestrin (2016) argued for. Interestingly enough interpretability, the ability to understand how a model works was not that important. Even though Hall & Gill (2019) argued that for a ML model to be trusted, it needs to be understandable as was brought up in section 2.3.2. Instead of understanding how a model works, people would trust the model simply because of their trust in other peoples and organizations. If a company with good reputation used a ML algorithm for decision making, that would be good enough for them to trust it. The same if a person they trusted would say something like “That is a great model, it works perfectly and gives us time to do other stuff”, Just as interviewee C said: “Personally it could be hard for me to understand machine learning models, and I’m not that interested in machine learning or AI for that matter as well. But for me, it could be a friend that is telling me that it is fine or that if an organization which i trust uses a model like that, I would be okay with it as well”.

However, what this actually means is that individually it is not that important to be able to understand and interpret a ML model. But for people to trust them, someone has to understand them just as Hall & Gill (2019), Ribeiro, Singh & Guestrin (2016) and Molnar (2020) argued for that was brought up in section 2.3.2. That means that since all interviewees thought that trust was a big factor, then indirectly interpretability and the ability to understand a model is just as important because somewhere along the line, someone has to understand
the model that is used. That knowledge creates a chainreaction between people and organizations who trust each other. The flow could be that one person understands the model, a second person trusts whatever the first person says and a third person trust what the second person says, and on it goes. But in the bottom of the chain, the model has to be interpretable and understandable.

5.3.2 Increased ability to work proactively

All of the interviewees mentioned several effects that could follow using a machine learning model to predict employee attrition. Almost everyone talked about how a recruitment process that aims at finding a replacement could be starting in earlier stages. Recruitment processes cost a lot of money (Dube, Freeman & Reich, 2010; Tziner & Birati, 1996; Zahra, 1993). And starting recruitment processes early to replace someone, can increase costs for the organization. Specially in the case where the person doesn’t quit who was supposed to quit according to the model. Today, a recruitment process often start when the employee tells the organization that he want to quit. With a model that can predict employee attrition, the process could potentially start much earlier (Tomassen, 2016). But the risk still remains, where you can end up with a recruitment process that has to be cancelled, or even worse end up with an abundance of employees. However, with starting recruitment processes earlier, organizations could minimize the notch in productivity that often occurs when replacing employees due to lack of knowledge and learning period.

We could also see that 6 interviewees said that a ML model to predict employee attrition could be used to investigate why people want to leave an what is the causes that make people leave. As interviewee C said in section 4.3.2: “If we can see with a model that people are leaving our organization, we have to ask the question why. With that information, we could start looking into factors that contribute to that and take action based on what the findings are. ”. This could help organizations and human resource units with their retention management. As was brought up by Kagmar et al. (2006) in section 2.1.1, one of the most important steps in reducing employee turnover is to understand the underlying reasons for the phenomenon. Successfully finding out underlying reasons to why people leave could grant important knowledge that can be used to fix problems which could minimize the chances that other employees leave as well. That would help organizations to retain their employees and therefore also their knowledge. Retaining employees would also decrease employee turnover. Lower employee turnover itself, as stated in section 2.1.1, has a positive effect on organizational effectiveness and competitive advantage (Koys, 2006; Davis, 2013; Morrow & McElroy, 2007; Dess & Shaw, 2001). It would also lower retraining costs, recall from section 2.1.1 that only in the U.S. fast food industry, $4.3 billion was annually paid in retraining costs (Kagmar et al. 2006). So organizations could benefit financially by lowering retraining costs, using information acquired from machine learning algorithms to predict employee attrition. Another effect is that the improved retention management and lowered employee turnover
can increase the organization’s attractiveness which increases the ability to attract and recruit new employees (James & Mathew, 2012). That, also affects the competitive advantage (Tomassen, 2016; Marjanovic, Cecez-Kecmanovic & Vidgen, 2018).

5.3.3 Retention management

The importance of retention management does not stop at lowering employee turnover (James & Mathew, 2012; Koys, 2006). Several interviewees witnessed that people leaving can break good relations and that it is important to keep good relations for the motivation among employees. As said above, a ML model to predict employee attrition could improve retention management and decrease the risk of employees leaving (James & Mathew, 2012). This would mean that there is a smaller chance that relations break and this will keep the motivation between employees. Motivation in turn is important for the performance of employees (Ganta, 2014). Ganta (2014) argues that an employees motivation is directly connected to the productivity, and that increased motivation also increases the productivity. An employee who is motivated is more likely to perform their task as good as they can, hence the increase in productivity (Ganta, 2014; Nothria, Groysberg & Lee, 2008). On the contrary, unmotivated employees are more likely to do deliver poor performance and there is a higher chance that they would leave the organization (Nothria, Groysberg & Lee, 2008). That in turn, could have not only direct impact on the performance, but also increase costs for an organization due to increased employee turnover.

As was shown in section 2.1.2 by James & Mathew (2012), retention management has a strong effect on the ability to maintain employees and maintaining employees will reduce employee turnover. Empirics showed that interviewees thought that information extracted from a machine learning model which could predict employee turnover could be used to work proactively by acquiring information. This means that the ability to understanding the underlying reasons to employee turnover would be increased. This was brought up as an essential part of reducing employee turnover in section 2.1.1. And that shows that using a machine learning model to predict employee attrition would have an effect on employee turnover and therefore also on an organization (Koys, 2006; Davis, 2013; Morrow & McElroy, 2007; Dess & Shaw, 2001).

5.3.4 Organizational culture

The empirics also state the organizational culture could change when people leaves. According to Mcgregor & Doshi (2015), the culture within an organization drive the motivation of employees and therefore also the success of the organization. A ML model to predict employee attrition could as stated above lower employee turnover and have positive effects on retention management. Therefore, it could also help an organization to preserve its culture and minimize the risk of culture changes that could decrease employee motivation.
The risk of cultural changes within an organization when employees quit is however low within a large organization, the smaller the organization the bigger the risk. Specially when units leave and not only single employees.

5.3.5 Decisions followed by a ML model
6 of the 8 persons who were fine with the usage of a ML model to predict employee attrition also said that the results should only be used as an indicator and that the decisions cannot be based solely on the model. As stated in section 2.2.1 by O’Neil (2016) and Marjanovic, Cecez-Kecmanovic & Vidgen (2018), algorithmic decision making in transformative services has potential to be disastrous. It is that fact, the potential consequences that also makes so that employees don’t want the algorithms to make the final decision. But the thing is, humans make disastrous decisions as well. And interviewee D explained further that computers are probably more unbiased than humans, and that there is a smaller chance that a decision made by a computer is unfair than when a human would make a decision in the same situation. The question here is whether the fact that not knowing exactly how a model works actually plays a bigger part than what interviewees initially thought of, since interpretability was not a big thing according to the empirics.

This means that using an ML model to predict employee attrition would be welcomed as a tool in a process. As was discussed earlier, it mattered more about trusting the model that is used. And the empirics in this study differed slightly from theory that was brought up in section 2.3.2 where Hall & Gill (2019), Ribeiro, Singh & Guestrin (2016) and Domingos (2012) argued that to trust a model you needed to understand it. However, interviewees in this study showed that personally understanding how a model works was not that important.

5.3.6 Transparency
The empirics stated that transparency was a key factor for usage of a model to predict employee attrition. 7 of the interviewees said that everyone that somehow is involved in a model like that should be aware of its usage. That is, employers need to be transparent towards their employees that a model is used, how it works and what the intentions are. Otherwise it could be used for the employers own benefits only. This is also strengthened by Farrell (2016) who argue that transparency is a key factor for a healthy relationship between employees and the employer. Not only does it affect the relationships, but can also strengthen employee engagement towards their organization. However, there was also found that the interviewees even though transparency should exist, the results of such a model should be handled by an unbiased unit within the organization, and the suggestion was HR. That is to minimize the risks of the usage against the employee instead of favouring them.
As was said above, the information about when individuals quit could be used for the employers best interests on the cost of the employee. As the interviewees said, if employers know that an employee will quit then could their behaviour towards that employee change. That could lead to consequences for both the employee and the organization. A prediction may be wrongly made and an employee who the model says is going to quit, might not quit. If the employer changes its behaviour based on the prediction, that could force the employee away from the organization. The effects would firstly be that the employee needs to change jobs, the organization will lose an employee which they would not have and therefore also lose the knowledge that was attached to that employee. Secondly, it could also have effects on the organizational reputation since people talk and a situation like that where behaviour towards an employee changes for the worse is something that would be communicated by the victim.

Also, adopting machine learning models to predict employee attrition could generate information about employees that can raise ethical questions and violate employee privacy (Bandara, Fernando & Akter, 2008). And raise questions such as who should be able to take part of the information, and what information is okay to use. And that is something interviewees discuss as well. However, Bandara, Fernando & Akter (2008) argues that if employees have a chance to participate and influence the analytical work that is done within HR, it would affect their perceptions about analytics and therefore increase their willingness to use them. And for that to be achieved, transparency is a key factor.

### 5.3.7 Algorithmic pollution

Side effects from decisions made from algorithms was brought up during interviews. This is known as algorithmic pollution from the theory section. Two interviewees said that they have a hard time believing that algorithms are completely unbiased and that groups of people might be affected because an algorithm favors other groups of people. Recall the example from section 2.2 where Amazon in 2018 specifically had removed gender from their ML model as a criterion, but even then the model learned in a way so that it was biased towards one specific group of people. The algorithmic pollution is a problem, but yet again we have to ask the question if humans make a better job? And as interviewee D said, the positive and the negatives when using a ML model for predicting employee attrition has to be weighted towards each other. If algorithmic pollution is on the negative side, then it has to be weighted towards the positive side and the decision to use the model can be made from that. And if then an organization accepts these facts, then they also have to accept that algorithmic pollution might exist.
6 Conclusion

The conclusion is based on the analysis and discussion of the empirics and theory to try to answer the two main research questions and the sub research question that could be found in section 1.4. The questions that was used were:

1. How does predictions with a machine learning model perform compared to a simpler regression analysis model when applied to HR data sets to predict employee attrition?

2. How does understanding a ML model, that predicts employee attrition, affect people’s will to use them?

2b. What do practitioners believe that such a model can contribute with to an organization?

From the analysis, it was possible to see that neither random forest, support vector machine or the multiple logistic regression model could be used to predict employee attrition with the data sets that were tested. The partitioning within the datasets that each model used was heavily balanced with employees who did not quit their job after a pulse survey and that could have affected the results negatively. Therefore, the machine learning models used, random forest and support vector machine, could not be compared in terms of performance with a multiple logistic regression model.

For the second research question, it was shown that it was not necessary for individual employees to understand how a machine learning model works in order for them to use them, or be used by them. Instead, the important part were that employees needs to be able to trust a model to accept it’s usage. To trust a model it was shown that it was enough that someone that the employees trusted in, trusted the model or that it was an organization with good reputation that used the model. Trusting a model was the most important part, but it was also shown that indirectly that meant that the models used has to be interpretable and understandable by someone who firstly can understand the model and secondly start a chainreaction of trust in the model. The second part of the will to use machine learning models were that decisions should not solely be based on the model, and a human should always be assisting decisions. Lastly, transparency in how the models are used and what they are used for was a key factor for employees to accept the usage of ML models to predict employee attrition.

For question 2b, it was shown that practitioners believed that the organization could be affected in several ways by being able to predict employee attrition. Arguments were made that such a model could be used to work proactively to improve retention management and therefore minimize employee turnover. It was shown that the effect that follows would have
an impact on organizational performance because the employee motivation would be maintained and knowledge that employees have would stay within the organization. That has direct impact on employee performance and therefore also organizational performance. Lowering employee turnover would also have direct positive impact on the economy of the organization. Lowering employee attrition would reduce costs in hiring processes and training new employees. This could also increase the attractiveness of the organization which could attract new talented employees.

Predicting employee attrition with a machine learning model could also contribute to that organizations could start recruitment processes earlier and therefore minimize the productivity notch when switching between old and new employees. This could help the organization to maintain the production on normal levels, and lower the chance of disruptions in productivity.

As a summary of the sub research question, it was possible to see that practitioners thought that a machine learning model which could predict employee attrition could decrease costs in recruitment processes, keep important knowledge within an organization, work proactively with retention management to lower employee turnover and to increase organizational attractiveness.

6.1 Future research
Recommendations for future research is to investigate further which models that could be suitable to find a model that is accurate enough to implement in the real world. Also, a deeper understanding of what variables and data that should be used when trying to predict employee attrition. Another interesting topic is to investigate if the effects of using such a model is positive or negative for the individual employee. Also, this technique could be applied to several other interesting possible predictions such as predicting sick leave, motivation, salary etc.
7 Bibliography


Linkedin. 2018. The Rise of Analytics in HR - The era of talent intelligence is here.


The R project. 2020. What is R?. URL: https://www.r-project.org/about.html Acquired: 2020-04-03


Acquired: 2020-04-28

Hastie, T.; Tibshirani, R. & Friedman, J. 2008. The Elements of Statistical Learning Data Mining, Inference, and Prediction. 2 ed. Springer International Publish


Acquired: 2020-05-04


McGrath, E. Joseph. & Brinberg, David. 1983. External validity and their research process: a


