Master's thesis

A quantitative research about predicting employee turnover for recruitment agencies through Artificial Intelligence.

IBRAHIM FISTIKÇI

STUDENT NUMBER: 2064307

MASTER INFORMATION MANAGEMENT

TILBURG UNIVERSITY

Company:

Kasparov Finance & BI

Thesis committee:

Ekaterini Ioannou

Ekaterini.Ioannou@tilburguniveristy.edu

Luca Heising

L.M.Heising@tilburguniversity.edu

June 9, 2022

Abstract

Many organizations struggle to attract and retain employees. With the power shifting to employees in a booming labor market, the pressure to of retention gets higher. This process is summed in the employee turnover. It is a KPI that measures workers who leave an organization compared to employees that are new or already working for the organization. It is used as a tool for management to examine the reason for the turnover percentage and act accordingly to decrease the outcome. This thesis addresses the integration of Artificial Intelligence within the recruitment sector by developing a model that predicts employee turnover. Four algorithms have been explored and the tested for its accuracy on the data provided. The method that is chosen in this research is a supervised machine learning algorithm: the decision tree. This algorithm is known for creating a simple, readable roadmap by splitting nodes in a yes/no format. It can be used to classify employee turnover: whether an employee left- or stayed within the organization. The model proved that it is able to accurately predict employee turnover with the condition that enough data is provided to train-, validate-, and test the model.

Foreword and acknowledgements

This research marks my last days as a student at Tilburg University. I would like to thank Ekaterini Ioannou for all the help and guidance during the entire process of my master's thesis. Also, I would like to Rens Verschuren, who made it possible to team up with Kasparov Finance & BI and write this research. Using Python to program a machine learning algorithm was a great way to challenge myself. I am very happy with the result and I hope that you will enjoy reading my master's thesis.

Ibrahim Fistikçi Tilburg, 09-06-2022

Table of Contents

List of figures

List of tables

Chapter 1

1.Introduction

1.1 Problem indication

> *"Robots will be able to do everything better than us… I am not sure exactly what to do about this. This is really the scariest problem to me." – Elon Musk* (Clifford, Catherine, 2017).

Robots are efficient and maintain a consistent level, whereas for people, the human error has to be accounted for that leads to a loss in consistency (Pratt & Murphy, 2012). Artificially intelligence robots are controlled by Artificial Intelligence (AI) algorithms (Gordon, 2020). Up until recently, robots could only be programmed to perform repetitive movements. Today, robots can be programmed intelligently to perform more complex tasks with the use of AI. The outcomes are future changing such as manufacturing robots, self-driving cars, automated financial investing, virtual booking agents, and the list goes on (Daley, 2021).

Deloitte defines AI as "getting computers to do tasks that would normally require human intelligence" (Duin & Bakshi, 2017). The purpose is to design machines with intelligence levels matching or exceeding the human brain (Brooks, 1991). Artificial Intelligence, and specifically, Machine Learning will play a leading role in improving operational efficiency and business decision-making. AI was firstly developed in 1950 by Alan Turing, a British polymath who explored the mathematical possibilities of AI (Buchanan, 2005). Turing suggested that available information can be used to make decision for people, so why can't a machine do the same thing? His 1950 paper discussed how to build- and test intelligent machines (The History of Artificial Intelligence, 2017). Intelligent machines should be used to solve problems and make decisions. Through the years, this concept developed significantly. More and more organizations recognize the benefits of AI and try to implement it to improve their decision making. One of the most popular methods in AI is machine learning (ML). It is a subset of AI and uses computers to simulate human learning activities by learning from previous events and optimizing its performance by developing intelligence (Bini, 2018).

According to Gartner Inc., the worldwide AI software market will reach $62.5 billion in 2022, an increase of 21.3% from 2021 (Rimol, 2021). According to the Fortune Business Insights, the global AI market value is expected to reach $267 billion by 2027 (Fortune Business Insights, 2021), and according to PwC, the total contribution of AI to the global economy is expected to hit $15.7 trillion by 2030 (PwC, sd).

1.2 Problem statement

There are countless ways to apply AI techniques for real-world problems. Many authors have researched this concept for many different industries globally. However, research on applying AI techniques in the recruitment sector is not exploited thoroughly yet, despite of the very competitive market and its effects.

Research shows that there is a significant shortage in finance- and IT professionals in the Netherlands (BNR Webredactie, 2021). According to Yacht, a big recruitment agency in the Netherlands, currently, there are four times more vacancies than employees in IT (Meijer, 2022). This factor makes the current market of finance- and IT professionals very competitive,

which leads to high pressure at recruitment agencies to attain them, and more importantly keeping them at their organization. According to van Vuuren, a professor from Tilburg University, The shortage of people in the labor market is at an extreme high. It is because of that many sectors opened their door at the same time after COVID-19 restrictions and are in need of workers again. Also, the economy is at a high level. Consumers dare to spend money again and therefore, there is a great need for labor (van Vuuren, 2022).

Companies usually offer higher wages to attract new professionals but still struggle to retain them because people consider more factors for their job than only their wages. (Krajcsák & Kozák, 2018). Employees working and leaving affects one of the most important key performance indicators (KPIs) within the recruitment sector: employee turnover. Employee turnover is a measurement that measures the duration of employees to stay at one organization and how often they have to be replaced. This ratio is called the turnover (Woods, sd). For recruitment agencies, revenue generated by its employees is their main source of income. This indicates that, if employees leave, it will directly affect the recruitment agency in question. Bearing in mind that the process of hiring a new employee can be very expensive, ideally an organization would like to retain their employees for as long as possible (Bowen & Ledford, 1991).

There is very little research on applying AI techniques on predicting the employee turnover within the recruitment sector. This topic should be researched because of the extreme shortages in the labor market and therefore, high importance of attracting and keeping employees, focusing on finance- and business intelligence (BI) employees.

This research will be conducted at Kasparov Finance & BI (also called Kasparov), a medium-sized recruitment agency focusing on fulfilling the (short-term) needs of clients on their finance and/or BI issues since 2010. This ranges from day-to-day financial tasks for a certain period, a CFO or CIO, or regular finance employees or BI consultants according to the needs of the client. Kasparov earns its revenue by freelancing its own interim pool to clients for an hourly rate. The main source of income for Kasparov are their employees. The interim team of Kasparov consists of a pool of employees that are generally early in their career with 0-5 years' experience in their work field. Working on short-term is attractive for young professionals since it allows you to work for different organizations in different functions (within finance and BI). It enables you to learn different cultures within different organizations and while doing so, build up a great resume.

Currently, Kasparov Finance & BI has no direct view on their employee turnover as a KPI. This means that, it does not have a good view of their most valuable asset, its employees. The organization's goal is to grow significantly over the years and therefore, they need to have a good overview of their employees leaving and new hirings in order to accomplish their goal. Therefore, this organization would need a tool to predict their highly important KPI in the future in order to account for the loss, opportunity costs, or by acting prematurely in order to keep the employee at the company instead of leaving. In order to tackle this issue, multiple machine learning techniques are going to be explored in order to choose the best suited model to predict the employee turnover for Kasparov.

1.3 Research questions

The research question to answer this problem is formulated as follows:

RQ: 'How can the employee turnover of a recruitment agency be predicted using Artificial Intelligence techniques?'

In order to know what AI is and which techniques can be used to a prediction analysis, the first sub question is:

1. *Which Artificial Intelligence techniques can be used to predict outcomes?*

It is also important to get good insights of which factors are most important when analyzing the employee turnover. Therefore, the second sub-question is:

2. *Which variable(s) cover the measure of employee turnover the best?*

The sub questions can be answered using the literature in chapter 2. However, the research question as well as the sub questions will be answered in chapter 5, discussion and conclusion.

1.4 Research design and -method

This study aims to develop a machine learning algorithm for management at Kasparov Finance & BI to predict its employee turnover. This study is designed to create new insights in using machine learning to predict turnover for any other recruitment agency. As mentioned in the problem statement, employee turnover paired with artificial intelligence is not yet explored thoroughly within the recruitment sector. Therefore, this research focuses on both 'artificial intelligence', and in specific, machine learning, and 'employee turnover'. So to answer the research- and sub questions, literature research will be conducted on the following topics to answer the formulated sub questions:

- *Artificial Intelligence*
- *Machine learning algorithms*
- *Prediction analysis*
- *Employee turnover*

The purpose is to gain a good understanding of the different machine learning techniques, how prediction analysis looks like, and which factors (or variables) are best suited to represent employee turnover, respectively.

After literature research, Kasparov is willing providing this research a dataset containing information from all current working employees and employees left within a period of three years. Considering that Kasparov is a medium-sized enterprise, providing only three years of information, the number of rows (employees) in the dataset will be relatively small. Building a machine learning model requires lots of data and therefore it will not be able to directly build a machine learning model. Therefore, an external dataset from the Internet will be retrieved that has a lot of similarities. Both datasets will be compared and transformed in chapter 3, methodology. The model will be build based on the data of the external dataset and later, Kasparov's dataset will be added and tested for its accuracy in chapter 4 results. The best machine learning algorithm will be chosen based on the size of the dataset, types of patterns, underlying assumptions, the level of noisiness, and the goal of the analysis (Shmueli, Bruce, & Patel, 2020). This will be discussed in chapter 5.

Chapter 2

2.Literature

This chapter reviews the concepts that are being used based on existing literature. It starts with discussing and reviewing the concept of AI, and in specific, machine learning. Also, rules of thumbs are reviewed concerning accuracy of a model and the size of dataset. Lastly, the concept of employee turnover paired with its important factors are being summed and reviewed.

Literature review

Search engines

This research has made use of the following databases to gather relevant information:

- Google Scholar
- Research gate
- Tilburg university libsearch
- Google search

Selection of literature

During this research criteria has been set for referencing:

- Literature regarding AI, ML, ML methods and -algorithms, rules of thumbs for accuracy and size of datasets, recruitment agencies, employee turnover and its affecting factors.
- Academic articles are carefully selected by citing from top journals

2.1 Artificial intelligence

There are tremendous volumes of data available in today's world. More than any human can process in its lifetime. In the last three years, 90 percent of the world's data has been created (Einstein, 2019). Today's data is beyond the bounds of singular human comprehension. Therefore, AI methods are established to assist organizations in their decision making. But when can we really say that a system constructed by a human is intelligent? According to Ertel, intelligence has two definitions (Ertel, 2018):

- 'Someone's intelligence is their ability to understand and learn things'
- 'Intelligence is the ability to think and understand instead of doing things by instinct or automatically'

According to Duin and Bakshi, AI is defined as 'getting computers to do tasks that would normally require human intelligence (Duin & Bakshi, 2017). There are many subsets of AI, but this research focuses on subset Machine Learning.

2.2 Machine learning

Machine learning (ML) is a subset of the broad concept of AI. This technique uses computers to simulate human learning activities by continuously improving performance and achieving self-improvement methods (Haoyong & Hengyao, 2011). In other words, it learns

from past experience and constantly learns and optimizes its performance by developing 'intelligence' over time (Bini, 2018).

ML has grown fast in the last years on data analysis and computations that allows to function intelligently (Sarker, 2021). It is quoted to be the most popular technologies in 'Industry 4.0' (Sarker, 2020). Industry 4.0 is the ongoing automation of practices such as data processing, using machine learning automation " (Sarker, 2020). Machine learning is the key to analyzing data and developing real-world applications. Machine learning algorithms can be categorized into four types: supervised, unsupervised, semi-supervised, and reinforcement learning (Shmueli, Bruce, & Patel, 2020).

Supervised learning

Supervised learning requires training data sets with known outcomes for the desired result, which means that the machine is taught by example (Shmueli, Bruce, & Patel, 2020). An existing dataset with the desired results beforehand is fed to the machine learning algorithm. Then the machine finds a method to arrive to those results while the maker already knows the correct answers to the problem. The algorithm identifies patterns in the data, learns from the observations, and makes its predictions. If needed, the maker corrects the wrong predictions of the model until it achieves a high level of accuracy (Shmueli, Bruce, & Patel, 2020). Examples of supervised learning are classification, regressions and forecasting.

Unsupervised learning

Unsupervised learning does not require any training dataset. There is no existing desired output or human operator to provide instructions to the machine. The machine analyzes available data to find relationships and correlations (SAS, 2022). In unsupervised learning, the ML algorithm interprets large datasets and addresses the data accordingly. The data gets categorized into clusters and as it assess more data, the ML algorithm improves more and more (Shmueli, Bruce, & Patel, 2020).

Semi-supervised

Semi-supervised learning is similar to supervised learning. It uses labelled and unlabeled data. Labelled data has relevant tags so that the algorithm can understand the data, while unlabeled data does not get relevant tags (Shmueli, Bruce, & Patel, 2020). This allows the machine learning algorithm to learn to label the unlabeled data as well.

Reinforcement learning

Reinforcement learning gets provided a set of actions, parameters, and end-values. This acts as defining rules for the algorithm. The ML algorithm then tries to delve into different options and possibilities by evaluating each result to then determine the best choice (Shmueli, Bruce, & Patel, 2020). By doing that, the machine teaches trial and error. By learning from past experience, it is able to adapt to the situation to achieve the best possible result (Wakefield, 2022).

2.2.1 Predictive analytics and intelligent decision-making

Machine learning is very popular because of its learning capability from the past and making intelligent decisions (Bini, 2018). This research focuses on predictive analytics and intelligent decision-making.

The application of data-driven predictive analytics to enable intelligent decision-making is very popular (Cao, 2020). It is all about capturing and exploiting relationships between explanatory variables and predicted variables from earlier events to predict unknown outcomes out of the future (Han, Kamber, & Pei, 2011). It can tackle many real-world problems like identifying criminal transactions such as credit card fraud. It could also serve as a tool for retailers to gain a better understanding in consumer behavior or inventory issues such as avoiding the risk to be out of stock or predict various human resources (HR) cases such as recruitment or hiring and firing.

2.2.2 Binary- and multi-class outcomes

One of the most essential steps of machine learning is to define the outcome, y, and decide how to measure it. Your desired outcome will dictate which different ML methods will be applicable for you. Hereupon, we distinguish between two types of classification: binary- and multi-class classification. A binary outcome can only take two values: 0 and 1, which often represents 'yes' and 'no' (Koyejo, Ravikumar, Natarajan, & Dhillon, 2014). For recruitment agencies this could answer the question whether an employee is working for them or not. Whereas a multi-class prediction can take two or more classes as an outcome (Aly, 2005). For example, if the employee is not working for the organization anymore it could answer for which reason they have left or if they are fired.

This research focuses on a multi-class classification in order to get better insights on how to create a well-functioning ML algorithm that could predict multiple outcomes which will be used later this research.

2.2.3 Clustering and classification

Clustering

Clustering is a popular technique used to discover sample groupings within data. Clustering is the process of grouping a set of objects into not predefined categories of similar objects. Objects in a group will be related to one another and different from the objects of all the other groups (Shmueli, Bruce, & Patel, 2020). Clustering is the most commonly used with unsupervised learning, which is discussed previously in this chapter.

Cluster analysis creates groups of objects, such that the objects within a group will be related between each other, and unrelated to the objects in other groups. The goal is to group together similar data, which defines the distance without a class label being available. The similarity measure is often more important than the actual clustering algorithm because it defines a distance function to cluster similar data points into same clusters such as the Euclidean- and Manhattan distance. It is important to validate clusters because the ultimate goal of clustering is to create meaningful clusters. It answers two important questions: are the resulting clusters valid, and do they really generate some insight (Shmueli, Bruce, & Patel, 2020)?

Classification

In contrast to clustering, classification assumes the existence of predefined classes. It trains a model that allows classifying new records to one of the classes. It tackles a predictive modeling problem where a class is predicted for a given example of input data (Shmueli, Bruce, & Patel, 2020). In contrast to clustering, classification is the most commonly used with supervised learning, which is discussed previously. Classification analysis identifies the

category of new observations on the basis of training data. The algorithm learns from a given dataset and classified new observations into a number of classes accordingly. There are two types of classification: binary classifier and multi-class classifier. Binary classifier exists when the classification issue has two possible outcomes. Think of a yes or no outcome. Whereas a multi-class classifier has more than two outcomes. Classification has many models that can be applied in various environments. Each model has its own pros and cons based on the type- and amount of the dataset available. Below are the most used classification algorithms used.

2.2.4 Machine learning algorithms for classification

Logistic regression

Logistic regression is a classification algorithm that is taken from the field of statistics for the application of machine learning. It uses statistics to analyze a dataset when there are one or more independent variables that determine an outcome. The goal of logistic regression is to find the best fitting model that describes the relationship between the dependent and independent variable.

It uses a logistic function to model the dependent variable. Logistic regression has three different types: (1) binary, in which it can only classify the dependent variable if the outcome is binary, (2) multinomial, for three or more unordered types, and (3) ordinal, for three or more possible ordered types. In the case of this research, the employee turnover contains a binary outcome; either the employee is working for the firm or he/she has left (0 or 1).

According to a study of sample size guidelines for logistic regression, a minimum sample size should be 500 records. Also, you could tailor it better to the research by use the following calculation: n = 100 + 50i, i refers to the number of independent variables in the model (Bujang, Sa'at, Abu Baka Sikdik, & Joo, 2018). This indicates that, this technique requires a rather small dataset compared to other techniques such as deep learning techniques.

K nearest neighbors

In k-nearest-neighbors (k-NN) the goal is to identify records (k) in the training dataset that look similar to a new record that we want to classify. This results in groups of similar records. These records are then classified into a class. And at last, the new record will be assigned to the closest neighboring class (Shmueli, Bruce, & Patel, 2020). The outcome could be either categorical or numerical.

This method does not make assumptions about the relationship between the variables. Instead, it shows the similarities between the independent variables in the dataset. An important factor is measuring the distance between the records based on the independent variables. The most popular measurement tool is Euclidean distance (Shmueli, Bruce, & Patel, 2020). This measure is widely used because it is computationally cheap. The Euclidean distance between records ($x1, x2, \ldots, xp$) and ($u1, u2, \ldots, up$) is:

$$\sqrt{(x_1 - u_1)^2 + (x_2 - u_2)^2 + \cdots + (x_p - u_p)^2}$$

Figure 1 - Formula Euclidean distance for kNN

After determining the distances between existing records and records to be classified, a rule is required to assign a class to the right destination. It is desirable to have k > 1. It should go as follows:

1. Find the nearest *k* neighbors to the record to be classified.
2. Use a majority decision rule to classify the record, where the record is classified as a member of the majority class of the *k* neighbors.

The advantage of choosing *k* > 1 is that higher values of *k* allow smoothing that will reduce risking to overfit. If k is too low, you risk to fit to the noise in the data. But if k is too high, you will not be able to optimally use the algorithm's ability to capture the local structure in the data, which is a big advantage of k-NN. Typically, values of *k* fall in the range of 1–20. It is desirable to use odd numbers to avoid ties. kNN works best with smaller datasets but struggles with larger datasets (Gupta, 2019). kNN stores all the data and then makes decision only at run time and therefore, it needs to calculate the distance of one given point with all other points. A large dataset cause a lot of processing which could affect the performance of the algorithm negatively (Kumar, 2019).

Decision tree

Decision tree separates records into subgroups by splitting on independent variables in the model (Shmueli, Bruce, & Patel, 2020).The splits create logical rules that are easy to understand. An example would be: IF Attending class <10 AND Education >5 THEN class = 1. A decision tree uses the independent variables of the dataset to create easy yes/no question and keeps splitting it until all data points belongs to each class (Bento, 2021). This creates a tree structure with all these yes/no options until the tree derives to the answer of the classification problem. These options are called nodes. So every time the tree asks another question, another node is added to the tree. The first node is called the root node. When you decide to stop the model after a split, the last nodes created are called leaf nodes.

Every split divides the dataset into the smallest subset possible. Therefore, the goal is to minimize the loss of function by splitting as little as possible . In order to measure how much data you have lost, we measure the impurity of the nodes. According to Shmueli, Bruce & Patel, two possible measures of impurity are Gini and Entropy. Both help to determine which split is best to build a pure decision tree (Shmueli, Bruce, & Patel, 2020).

$$\text{entropy(node)} = -\sum_{k=1}^{m} p_k \log_2(p_k)$$

Figure 3 - Formula Entropy Impurity

$$\text{GINI(node)} = 1 - \sum_{k=1}^{m} (p_k)^2$$

Figure 2 - Formula Gini Impurity

The decision tree has turned out to be a very popular classification technique in machine learning. It performs well across a wide range of situations and does not require much effort from the analyst. When the trees are not too large, it is easy understandable by the consumers. The classification tree is typically a good performer for small datasets. Using a decision tree works best when the dataset is small and simplicity is central in interpreting the data (Shmueli, Bruce, & Patel, 2020).

Random forest

Random forest is a classification algorithm very similar to the decision tree. It consists of a given number of individual decision trees that operate all at once. Each tree makes his own class prediction and the class with the most votes becomes the model's prediction. For example, if you use 5 decision trees, from which three says 0 and two says 1, the outcome is 0.

This algorithm's strength is also in its ability to create uncorrelated trees operating as a whole. Having low correlation between the individual trees can predict more accurately than any of the individual predictions (Yiu, 2019). Having many individual trees gives reduces the error because while having some wrong outcome trees, there will be many that are right so it will move to the desired direction.

Because it uses multiple decision trees, its accuracy tend to be higher than the decision tree, however, it makes it also more complicated to read, in contrast to the easy understandable decision tree. However logically, it also requires more data since it uses multiple decision trees (Zakariah, 2014).

2.3 Training-, validating-, and testing sets

Building a machine learning algorithm includes three types of datasets: the training-, validation-, and testing set. All three have their own contribution to a well-build machine learning model: training, tuning, model selection and testing. Such model has parameters and hyperparameters. Parameters are a configuration variable internal to the model from which the value can be estimated from the data (Brownlee, 2017). They are the values the algorithm can change independently as it learns. An example of a parameter is the coefficients in linear regression. In contrast to parameters, hyperparameters are a configuration external to the model and from which the value cannot be estimated from the data (Brownlee, 2017). They are specified by the researcher and is used to affect the parameters how we would like to refine the model. Examples of such are de k in k-NN or the k in k-means clustering. Below, the three sets are briefly explained.

Training set

The training set is the dataset that we use to train the model. It is able to see and learn the best possible combinations which will generate a strong predictive machine learning model. The goal of the training set is to create a trained model that generalizes well to new data (Brownlee, 2017).

Validation set

The validation set is a set of data used to train the machine learning model with the purpose of choosing the best model and optimizing it. Machine learning engineers use validation to refine the model's hyperparameters in order to indirectly affect the model (TechTarget, 2022). Overfitting is also detected and prevented to eliminate errors for future predictions and observations if an analysis corresponds too precisely to a specific dataset. Overfitting occurs when the new data points is too closely aligned to the observations.

Testing set

The testing set is used when the model is completely trained by the training- and the validation set(s). It is used to assess the performance of the machine learning model. The final model is used to test results and assess the performance of the final model (TechTarget, 2022).

2.4 Rule of thumb – Accuracy

Accuracy represents the accurate predictions in classifying tasks. It is calculated by dividing the total number of correct predictions to the predictions generated by the machine learning model (Barkved, 2022). According to an article in Towards Data Science, a data scientist claims that in the real-world, companies expect a minimum of 90% accuracy on the machine learning model (Shin, 2020).

2.5 Rule of thumb – Size of dataset

The size of a dataset refers to how many rows of data a dataset contains. In the case of this research, every row should represent an employee with features (or variables) as columns. Building a machine learning algorithm requires a certain amount of data but how much data is sufficient? This question is hard to answer because every dataset has its own characteristics as well as every machine learning algorithm. However, according to an article from the journal of choice modelling, a rule of thumb is that the dataset needs to be at least a factor 10 to 100 times the number of variables (Alwosheel, van Cranenburgh, & Chorus, 2018).

2.6 Employee turnover

2.6.1 Recruitment sector

The recruitment sector is a sales driven industry known for its high competitiveness and its fast pace. It plays an important role in responding to skill shortages and labor demands (Hickey, 2020). The Dutch temporary employment sector has grown by 15% in 2021, and a 20% growth for the administrative sector (Staffing Industry, 2022). But despite the high growth percentages, the Netherlands reported to have more vacancies than employees. For every 100 employees, 126 job vacancies are available (Victoria Séveno, 2021). Due to COVID-19, the unemployment rate increased even more, which increases the pressure on recruitment agencies even more. The demand for external recruitment increased due to skill shortages in the labor market.

2.6.2 Employee turnover rate

As mentioned before, employee turnover is a measurement of workers who leave an organization and are replaced by new employees (Woods, sd). It is used as a tool for management to examine the reasons for the turnover percentage and act accordingly to decrease the outcome. This means that the higher the turnover rate gets, the worse it gets for an organization. A high turnover can result in low employee morale because of long working hours, high responsibilities, and perceived salary (Dwesini & Sisulu, 2019). Whereas low employee turnover increases morale and increases overall productivity.

To calculate the employee turnover rate, you divide the employees who left by the average number of employees and multiply this by 100 in order to get a percentage as outcome (Price, 1977). See below the formula:

$$\text{Employee Turnover Rate} = \frac{\text{Employees Who Left}}{\text{Average \# of Employees}} \times 100$$

Figure 4 - Formula Employee Turnover Rate

2.6.3 Factors influencing employee turnover

Salary

Previous research has found employee compensation to be an important predictor of employee turnover (Lee & Whitford, 2008). Under the income tax act, salary refers to the compensation received by an employee from an employer for the execution of services in connection with employment (India Filings, 2022). Employees prefer their work efforts to be recognized and rewarded. However, organizations focus more often on their own revenues rather than its employees (Gregory, 2011). By rewarding employees monetarily, they would feel that their hard work is being appreciated, which they need (Branham, 2005). Through rewarding, employees tend to stay working for their current organization (Gregory, 2011).

Secondary benefits

Secondary benefits are rewards that are agreed upon with your employer on top of your salary (Consultancy, 2022). Secondary benefits are not necessarily financial. Examples of such benefits are a company car, a phone, receiving discount for sports facilities, or a performance-based bonus. According to the Chicago school of professional psychology, employee's job satisfaction and organizational retention rate increases with a compensation plan with room for bonuses and periodic pay rises (Determinants of Job Satisfaction in the Workplace, 2022)

Training & development

Aguinis and Kraiger thinks that the importance of training to improve organizational effectiveness for individuals and teams is high by developing their knowledge and skills (Kraiger & Aguinis, 2009). Another research discusses the importance of organizations recognizing that employees make a major contribution to the success of the company, specifically individuals who are highly motivated and considered to be top performers (Hollman & Abassi, 2000). This concludes that, organizations that fail to invest in investing in the development of its employees may fail to retain high performers and thereby risk losing extra revenue.

According to Taylor, employees who are given the opportunity to develop in their career within their current organization will be less likely to seek progression opportunities externally (Taylor, 2014). Two other researches support this statement by claiming that when clear internal career paths are recognized and professional development opportunities are handed out, it is one of the key factors contributing to employees staying or leaving a firm (Ahmad & Daud, 2016); (Rothwell, 2016).

Age & gender

Results of an empirical study on certified public accountants (CPAs) indicates that age is a significant moderator in the relationship between performance and their intention to quit (Werbel & Bedeian, 1989). This suggests that age, being a personal characteristic, is important to consider in effecting employee turnover (Vardi, 1980). The empirical study conducted by

Werbel & Bedeian (1989) resulted in a relationship between performance for younger and older employees. They suggest that older, poorer performers tend to stay at their organization longer than younger employees, whom tend to leave quicker (Werbel & Bedeian, 1989).

Vardi (1980) also suggests that gender, which is also a personal characteristic, plays an important role in determining the employee's work behavior and thus, the employee turnover (Vardi, 1980).

Size of employer

According to a research from Columbia University, the employer size and employee turnover have great impact on each other (Idson, 1993). The paper investigates the causes of low turnover for employees at large organizations. This research concluded that employer size affects labor mobility, which refers to how easy employees move around within an economy, which is strongly affects the employee turnover. It comes down to that large employers have more capabilities to establish a good relationship with its employees rather than medium or small companies (Idson, 1993).

This research categorizes the size of an organization in three different categories (European Commission, 2022):

- Small enterprise (1-49 employees)
- Medium-sized enterprise (50-249 employees)
- Multinational enterprise (>250 employees)

Satisfaction level

Job satisfaction is simply measures if you are satisfied with the job you are performing, whether the company is a match, and also involving personal intuition and circumstances (Bourne, 2022). The satisfaction level could be related to any of the factors named in this sub-chapter. According to research conducted on the job satisfaction and employee turnover relationship among nurses, it shows that job satisfaction was significantly associated with the turnover. Nurses who were unsatisfied with their job were 2.55 times more likely to leave than nurses that were satisfied. This research concluded that continuous effort should be made by managers to enhance job satisfaction (Gebregziabher, Berhanie, Berihu, Belstie, & Teklay, 2020).

Promotion

Getting promotion at work means that you will elevate in function and from for example an assistant accountant to an accountant. According to a report from ADP Research Institute that takes a close look at how pay and promotions relate to retention, overall, firms promote 8.9 percent of employees (Yildirmaz, Ryan, & Nezaj, 2019). It concludes that employees who get passed over for promotion while another employee advances, they are more likely to quit. It claims that "employers should consider the effect of promoting individual team members and the retention strategy for key team players" (Kuehner-Hebert, 2019).

Last evaluation

Frequent evaluations provides clarity to employees about the expectations of its employer. It is a platform that evaluates and values the performances of the individual (Dawson, 2021). New goals can be set for the sake of the individual as well as for the company to create a win-win situation. Also, it is a great tool to distinguish top-performers and low-performers and act accordingly in these situations.

2.7 Conclusion literature

Artificial intelligence refers to getting computers to do tasks that would normally require human intelligence and is used as an umbrella term for multiple possible techniques (Duin & Bakshi, 2017). This research focuses on subset machine learning. Machine learning can be categorized in four types: supervised-, unsupervised-, semi-supervised-, and reinforcement learning (Shmueli, Bruce, & Patel, 2020). It could also be divided in either a clustering method or classification method. This research will use a supervised machine learning classification because it can provide training data sets with known outcomes for the desired result and it assumes the existence of predefined classes.

The classification algorithms that have been investigated by describing the algorithms extensively are: logistic regression, k nearest neighbors, decision tree, and random forest. The reason for these algorithms is because of the relatively low requirement of dataset size that is described in each paragraph. Also, rules of thumb concerning accuracy of a machine learning model and the size of a dataset is reasoned (Alwosheel, van Cranenburgh, & Chorus, 2018).

Lastly, the employee turnover is explained. Employee turnover can be used as a key performance indicator. It is used as a tool for management to examine the reasons for the turnover percentage and act accordingly to decrease the outcome (Woods, sd). Literature finds that the following factors are highly likely to affect the outcome of employee turnover: salary, secondary benefits, training & development, age & gender, size of employer, satisfaction level, promotion, and last evaluation. The next chapter will identify which variables can be used during this research because of several reasons.

Chapter 3

3.Methodology

This chapter focuses on the important steps in the research process, starting with a step-by-step application of Pfeffer's design science methodology on this matter. This will illustrate the red-line throughout this research. After that, the focus switches to the data being used. First, the data collection process, overlapping data, and data transformation. Later, the machine learning model is being selected. Lastly, the chosen model will be explained in detail, focusing on Kasparov Finance & BI.

3.1 Design science

Design science is defined as: "research that invents a new purposeful artefact to address a generalized type of problem and evaluates its utility for solving problems of that type" (Venable & Baskerville, 2012). This research applies design science, and according to Peffers, design science includes six steps to successfully perform design science (Peffers, 2006):

1. Problem identification and motivation

Step one defines the research problem and justifies the value of the solution. This research aims to motivate the researcher and the audience of the research to pursue and accept the solution. This step is explored in detail in chapter one, paragraphs problem indication and problem statement.

2. Objectives of a solution

Objectives of a solution can either be quantitative where the solution would be better than current ones, or qualitative where a new artifact is created to solve a problem. This research applies a quantitative approach by using approaches and methods from previous literature in different situations and apply all the useful conclusions of those previous researches to create a better, upgraded version of a machine learning algorithm. This literature can be found in chapter two of this research, where AI techniques, and in particular, machine learning techniques, and employee turnover is explored.

3. Design and development

Step three includes the design of the artifact. Here, the architecture functionality is determined. The artifact designed in this research is association rules machine learning algorithm, predicting employee turnover using two datasets from the recruitment sector.

4. Demonstration

Step four demonstrates the efficacy of the artifact. This will be done by splitting the data described in the next paragraphs into training-, validating-, and testing sets for the ML algorithm. training set is subject to training the model to allow classifying new records to a certain class. The validating set to fine-tune the model, and a testing set to test whether the model predicts what is desired.

5. Evaluation

The fifth step starts by comparing the objectives in step two and observed outcome from step four. It will be evaluated in the chapter five, discussion according to the results in chapter four.

6. Communication

The last step concerns the communication of the entire research, from problem identification to evaluation will be communicated with everyone interest in this research by publishing it publicly.

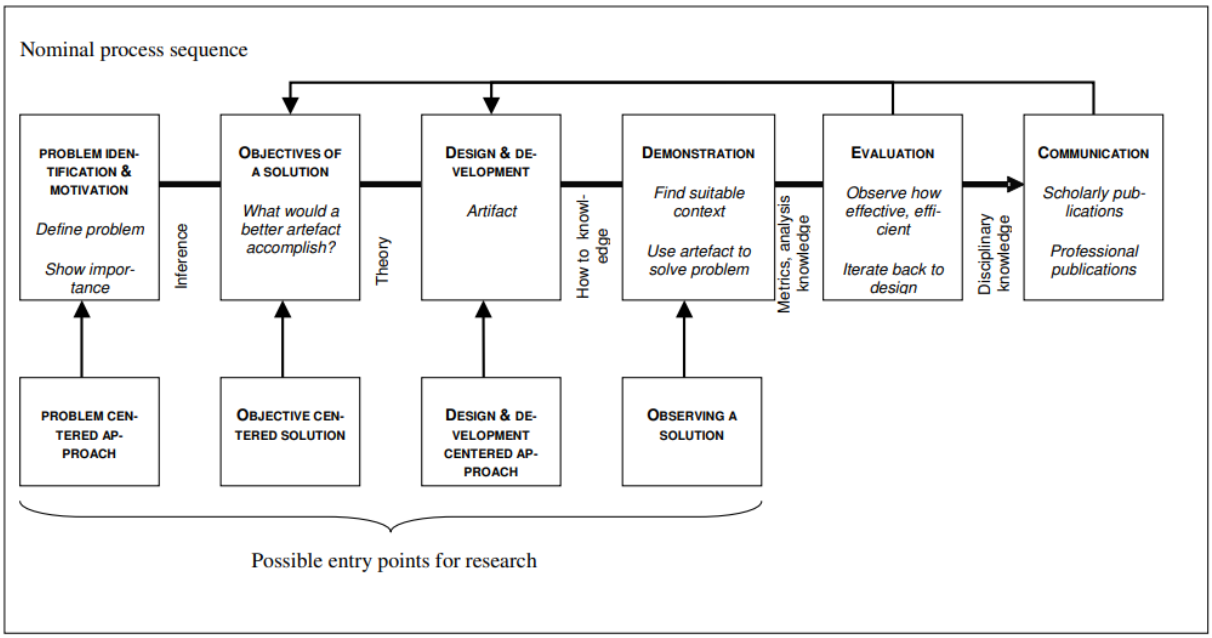


Figure 5 - Design science research process model (Peffers, 2006)

3.2 Data collection

This research collected data from two different sources: Kasparov Finance & BI and an external dataset online available on the Internet. Using two datasets is necessary because Kasparov's dataset is too small to build a machine learning model on. It contains a total of 96 rows of data, with each row representing a (former) employee. Building a supervised machine learning model requires more data since the dataset has to be split in a training set, validation set(s), and a testing set (Alwosheel, van Cranenburgh, & Chorus, 2018). A rule of thumb to determine the size of a dataset to build a model is to have 10 to 100 times the number of variables that is being used (Alwosheel, van Cranenburgh, & Chorus, 2018). This indicates that, since Kasparov's dataset only contains 96 rows of data, it cannot even contain one variable, so it is not possible to build a model solely on this dataset. Therefore, a second, external dataset has been retrieved containing 15,000 rows of data, which is going to be used because of its size and the high degree of similarity with Kasparov's dataset in terms of variables (or features) that was collected (GitHub, 2021).

Because two datasets are going to be used, they both have their own purpose within this research. The external dataset will be used to train, validate, and test the model. This will result in a machine learning model. Later, 50% of Kasparov's dataset will be added to the existing training data to fit the model and the other 50% to test the model. This should firstly result in a

machine learning model that predicts employee turnover based on certain variables (or features). And later be tested whether it also works for Kasparov specifically.

The data retrieved from Kasparov is collected and extracted from its external human resource partner called AFAS. The collected data consists of all current- and former interim finance or BI employees between 01-01-2019 and 01-03-2022. It consists of 96 rows of data and originally containing the following variables:

- Employee number
- Gender
- Time spend company (number of years spent in the company)
- Department
- Average monthly hours
- Home address
- Age
- Promotion last year
- Number projects (number of projects/assignments completed while at work)
- Salary (low, medium, or high)
- Left (whether the employee left the company or not)

The external dataset is retrieved from the Internet. It originally contained 15,000 lines, each resembling an employee from a multinational enterprise. The dataset is focused on employee turnover of a multinational enterprise. The data set originally contained the following variables:

- Satisfaction level (job satisfaction)
- Last evaluation (time since the last evaluation in years)
- Number project (number of projects/assignments completed while at work)
- Average monthly hours
- Time spend company (number of years spent in the company)
- Work accident (whether any work accidents have occurred for the employee)
- Left (whether the employee left the company or not)
- Promotion last year (whether the employee got promoted within the last year)
- Department (in which department the employee works)
- Salary (low, medium, or high)

3.3 Data overlapping & data transformation

As previously mentioned, both datasets should be similar in order to generalize the results of this research. The listed variables for both datasets have a lot of similarities such as: (1) number project, (2) time spend company, (3) average monthly hours, (4) department, which will be discussed later, (5) salary, and of course the dependent variable "left". It is important to note that it is not possible to add any variables to the external dataset since we have no influence in that. Whereas, for Kasparov's dataset, the researcher can request additional information if that information is captured by Kasparov. Therefore, we already have to exclude the following variables from Kasparov's dataset despite the relevance found by literature described in chapter 2.6.3: (1) employee number, (2) gender, (3) home address, and (4) age.

Having that said, the variable "work accident" will be excluded from the external dataset. This because of work accidents have yet never occurred in any way possible from Kasparov and therefore, will have little to no effect on the machine learning model on Kasparov's behalf. Also, the external dataset contained of ten departments. Many of which were irrelevant for Kasparov since they do not have those departments within its firm, such as product management and technical service. Therefore, only the four departments that are present within the organization will be included: finance, management, BI, and HR.

Kasparov does not collect the satisfaction level and the last evaluation so it is not possible to include this in the final model for Kasparov. However, literature in chapter 2.6.3 shows that the variables are important. Therefore, the model in the external dataset will include variables: satisfaction level and last evaluation, which will be written in brackets below, and Kasparov's dataset will not (Bourne, 2022). See below the variables:

- (satisfaction level)
- (last evaluation)
- Number project
- Average monthly hours
- Time spend company
- Promotion last year
- Department
- Salary
- Left (y)

This leaves the external dataset with 3,363 rows of data. The rule of thumb regarding the size of the dataset is that it needs to be at least a factor 10 to 100 times the number of variables. Having eight independent variables, the minimum size of the dataset should be between 800 and 8000 lines (Alwosheel, van Cranenburgh, & Chorus, 2018). This implies that, the external dataset contains enough data to build a machine learning model. Because no variable that has effect on the number of employees has been removed for Kasparov's dataset, it still contains 96 rows of data.

Appendix A shows the distributions for all variables within both datasets. Similar variables will be illustrated next to each other to be compared easily. Variables satisfaction level and evaluation time will be illustrated individually because it only exists in the external dataset. Differences such as in the number of projects, average monthly hours, and promotion last year is because of the dominance of the external dataset. Even though both datasets are recruitment based, it is hard to have exactly the same data points for 3,363 row dataset and a 96 row dataset.

In order to scale the numeric variables and encode the categorical variables in both datasets, a custom preprocessing pipeline is written in Python. Python is a programming language that will be used to build the machine learning algorithm. The newest version (3.10.0) is being used for all the codes concerning this model (Python, 2021). It plays a useful role in transforming and manipulating the data. All numeric variables are first imputed and then scaled. After that, all categorical variable are imputed and after, one-hot-encoded. One-hot-encoding means that for a yes/no answer, it changes it to 0 and 1, so Python can read it. It used the following sklearn packages:

```
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.impute import SimpleImputer
from sklearn.pipeline import Pipeline, make_pipeline
from sklearn.compose import ColumnTransformer, make_column_transformer
```

Figure 6 - Python code for preprocessing

After one-hot-encoding, feature names change because of the 0-1 split into many columns. Therefore, everything was renamed using the custom pipeline that was mentioned before.

3.4 Selecting the best model

After having two very similar datasets that is thoroughly explained in section 3.3, the best machine learning algorithm must be selected. The term ‘best machine learning algorithm’ should be interpreted as the machine learning algorithm scoring the highest accuracy. Accuracy represents the accurate predictions in classifying tasks. The outcome is expressed in a percentage showing the ratio between the total number of correct prediction to the predictions generated by the machine learning model (Barkved, 2022). Since the goal of this research is to build a machine learning model that will be used in practice by Kasparov Finance & BI, this model should be as accurate as possible. As explained in section 3.2, one of the purposes of the external dataset is to build the model and therefore, the machine learning algorithms are going to be tested on its accuracy based on the external dataset. In total, four algorithms have been tested for its accuracy: logistic regression, K nearest neighbors, decision tree, and random forest. These four algorithms have been included in this research due to the knowledge and familiarity of the methodology of each algorithm by the researcher. The process starts with importing the four algorithms in Python, assuming that the data transformation is finished.

```
# KNN, DecisionTree, RandomForest, LogisticRegression
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
```

Figure 7 - Python code for testing algorithms

In order to select the best fit for this dataset, 5-fold cross-validation is performed on each of the four algorithms. Cross-validation measures the ability of a machine learning model to estimate unseen data (Brownlee, 2018). This means that the dataset will be divided into five completely random, equal pieces. Every piece will be tested individually, and based on the accuracy percentages of these validation sets, cross-validation determines the accuracy of the machine learning algorithm based on the external dataset. Cross-validation usually is the preferred method because the model can train on multiple train-test splits instead of only one. This allows the model to determine a better indication of how well the model will perform on unseen data (Allibhai, 2018). Below, you can find the Python code for the cross validation and the accuracy scores for the external dataset.

```
lr_score = mean(cross_val_score(logistic_regression, X_train, y_train, cv=5, scoring="accuracy"))
knn_score = mean(cross_val_score(knn_regression, X_train, y_train, cv=5, scoring="accuracy"))
dt_score = mean(cross_val_score(decision_tree, X_train, y_train, cv=5, scoring="accuracy"))
rf_score = mean(cross_val_score(random_forest, X_train, y_train, cv=5, scoring="accuracy"))
```

Figure 8 - Python for five-fold cross-validation

```
LogisticRegression: 0.8053143171460003
K Nearest Neighbors: 0.9639187490177589
DecisionTree: 0.9785871444287285
RandomForestClassifer: 0.9888983184032689
```

Figure 9 - Python outcome of accuracy for each algorithm

Given the accuracy percentages in figure 9, random forest has the highest accuracy, closely followed by decision trees and K nearest neighbors. Logistic regression used for classification comes last with a significantly lower accuracy than the other three. As discussed, this research aims to build a machine learning model with the highest possible accuracy paired with the simplicity of the model. Logically, the interesting algorithms would be either decision tree, or random forest given they are the two highest scoring for accuracy. As explained in chapter 2, random forest is simply multiple decision trees combined, creating high accuracy. This comes at a cost of simplicity of the model. the decision tree is easy understandable for consumers and given that this model has to be understood not only by management as well as lower hierarchies in the organization such as HR, easy understandability is important and therefore, it is chosen to be used in this research.

3.5 Decision tree implementation

The next step in this research is to apply the decision tree algorithm on the data. This research divides the implementation of a decision tree in three steps: (1) mapping the important features, (2) deciding the number of splits, (3) displaying the results of all predictions by the decision tree in a confusion matrix, (4) and lastly, plotting and explaining the decision tree.

3.5.1 Important features

The most important features (or variables) refer to the features that will affect the model its predictions the most. These variables are most important and will also come forward in the decision tree itself. The first node in the tree is called the root node (Shmueli, Bruce, & Patel, 2020). This node always starts with the most important feature according to the data provided. The tree will continually split until the stop has been decided by the researchers according to a splitting measurement tool. However, the decision tree does not show how much impact the feature will have in comparison to others. This indicates that, it is only possible to see which

variable has more impact within the model but not the ratio. Therefore, a feature importance code in Python, called XGBoost classifier, will be executed in order to map the most important features in a good overview. Figure 10 shows the Python code for executing the feature importance.

```
from xgboost import XGBClassifier

# define the model
model = custom_pipeline(X, XGBClassifier()).fit(X, y)
# get importance
importance = model['classifier'].feature_importances_
# summarize feature importance
importances = {}
for i,v in enumerate(importance):
  importances[feature_names[i]] = v

importances = dict(sorted(importances.items(), key=lambda item: item[1]))
# plot feature importance
plt.barh(range(len(importances)), list(importances.values()), tick_label=list(importances.keys()), align='center')
plt.show()
```

Figure 10 - Python code for showcasing feature importance

3.5.2 Splitting decision

The second step is to decide the parameters of the decision tree; which are the number of splits of nodes. A decision tree makes decision by splitting nodes into sub-nodes. Its goal is to divide the data into smaller, homogeneous groups (Hssina, Merbouha, Ezzikouri, & Erritali, 2014). Homogeneity means that most of the data points at each node are from one class (Breiman, Friedman, Olshen, & Stone, 1984). After every split the model loses data. Losing data leads to inaccurate predictions within the model. The goal is to minimize the loss function as much as possible to build a pure decision tree (Bento, 2021). Gini impurity and Entropy impurity are two well-known measurements that compares the data distributions before and after the splits made to measure the number of data lost.

Both Gini-, and Entropy impurity measure the purity of the tree, which refers to whether a sub-node contains only data points of one class. For example, if you split root node “gender” to “woman” and “men”, and all the predictions “men” goes to sub-node “men” and the same for woman, it is a pure tree. However, in the real-world, data is not always classified in the right subset after a split and therefore, loses its purity. Gini impurity has a minimum value of 0, which is the best you can get, and a maximum of 0.5, which is the worst you can get (Shmueli, Bruce, & Patel, 2020). Entropy has a minimum value of 0, which is the best possible, and a maximum value of 1, which is the worst outcome possible (Shmueli, Bruce, & Patel, 2020).

This research aims to simplify the interpretation by plotting a graph that will exactly show the accuracy percentage of the machine learning model after any number of split. This percentage represents the accurate predictions in classifying tasks (Barkved, 2022). Based on that, and bearing in mind the need for simplicity, the number of splits will be determined.

3.5.3 Confusion matrix

A confusion matrix is a performance measurement for machine learning classifications (Shmueli, Bruce, & Patel, 2020). This matrix is used to predict the accuracy of the machine learning model. Figure XX shows a confusion matrix. It contains four squares being (Narkhede, 2018):

- true positive (**TP**) on the top left, which simply means that the prediction is positive and it's true
- false positive (**FP**) on the top right, which refers to that the prediction is negative but it is actually true
- false negative (**FN**) on bottom left, meaning that the prediction is positive but the actual is negative
- true negative (**TN**), meaning that the prediction is negative as well as the actual

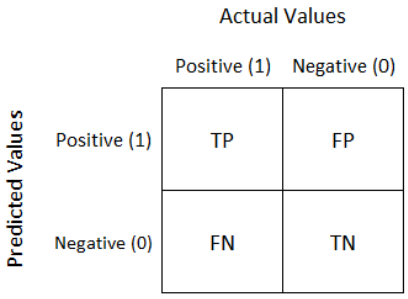


Figure 11 - General confusion matrix

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Number\ of\ all\ predictions}$$

Figure 12 - General formula for calculating accuracy

Figure 12 shows the formula to calculate the accuracy according to the confusion matrix. It calculates the total correctly classified data points divided by the total number of data points (Shmueli, Bruce, & Patel, 2020). This research aims to deliver a machine learning based model that predicts the employee turnover and therefore, having high accuracy is very important.

3.5.4 Decision tree

Reading and interpreting a decision tree is simple. It creates useful splits by separating records of data into subgroups. These splits create useful classification rules in an easy understandable yes/no format. Starting from the root node, you go do to the next nodes until you reach the leaf node, which tells you the outcome. In this case, the outcome would be 0 or 1, being stayed- or left the organization, respectively. How to interpret the results of this research will be explained in detail in the next chapter: results.

3.6 Conclusion methodology

Due to lack of data on behalf of Kasparov, a second, external dataset is going to be used to build a properly working machine learning model that predicts the employee turnover using classification techniques. Based on that model, Kasparov's dataset will be tested to verify whether the model also predicts Kasparov's data accurately. Both datasets are very similar in features, but the external dataset has two more features that is found important by supported academic literature: satisfaction level and last evaluation. Therefore, these two variables have to be kept out the final model that will be used by Kasparov.

The model that is going to be used in this research is the decision tree because of its high accuracy on the given data and its simplicity. Simplicity is very important for the stakeholders because the machine learning model has to be understood by higher management as well as regular human resource employees.

Chapter 4

4.Results

The machine learning algorithm that is applied in this research is the decision tree. This due to the simplicity in use paired with the 97.86% accuracy it projects based on the external dataset before determining the number of splits. According to an article in Towards Data Science, a data scientist claims that in the real-world, companies expect a minimum of 90% accuracy on the machine learning model (Shin, 2020). This indicates that, the current accuracy level of the decision tree is aligned with data scientists' expectations.

The results are divided in two main parts: (1) the results from the external dataset, and (2) the results of Kasparov's dataset. The layout is the same for both parts:

(1) mapping the important features,

(2) deciding the number of splits,

(3) displaying the results of all predictions by the decision tree in a confusion matrix,

(4) plotting and explaining the decision tree.

4.1 External dataset results

4.1.1 Feature importance

Figure 13 plots a bar chart showing the most important features (also called variables) based on the external dataset. It is scaled between 0 and 1, which can also be expressed in percentages. Therefore, the features are equally divided and a total will add up to 1. The chart shows the important features in descending order and therefore, the top feature is determined to be most impactful in the classification of the outcome.

This indicates that, satisfaction level is determined to be most impactful within this model. Followed by number project, time spend company, last evaluation, average monthly hours, and salary high, in descending order from high to low importance. All other variables show low importance, and therefore low impact on the data. This implies that, apart from the impactful features, the other variables have low to no impact on classifying the outcome for employee turnover.

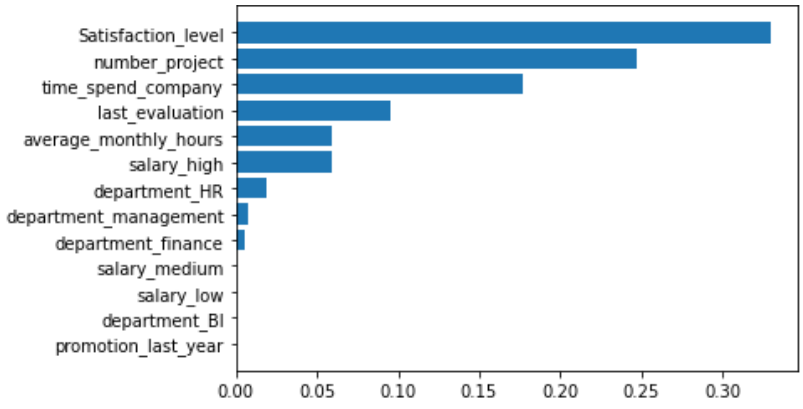


Figure 13 - Bar chart showing the most important features for the external dataset

4.1.2 Deciding the best split

Splitting nodes into sub-nodes divides the data into smaller, homogeneous groups with the goal to allocate the data points to one class (Hssina, Merbouha, Ezzikouri, & Erritali, 2014). The goal of this process is to have strong yes/no decisions that actually predict what is desired bearing in mind to keep the model as simple as possible. Simplicity is reached when the number of splits is kept as low as possible paired with high accuracy. If this process is not applied, the sub-nodes will not be reliable, which implies that the model is unreliable and thus, useless in the real-world (Hssina, Merbouha, Ezzikouri, & Erritali, 2014). The downside of splitting is that after every split, the model loses data, which leads to inaccurate predictions within the model. Therefore, the goal is to minimize the loss of data by building a 'pure' decision tree (Bento, 2021). Also, after every split, the model its simplicity is getting lower.

Two popular measurements to calculate the purity of the tree are Gini impurity and Entropy impurity. It compares the data distributions before and after each spit. By doing that, it can measure how much data is lost during the process and can accurately project this. Chapter 3.5.2 explains that both measurements have different scalers and therefore, the results are interpreted differently. To normalize the interpretation of the results for Gini and Entropy, this research plots a heat map where both outcomes are on the same scaler and can therefore be interpreted and compared easily.

Figure 14 shows a heat map showcasing the accuracy for Gini and Entropy after every split. The x-axis shows the number of splits, the left y-axis shows the method and the right y-axis shows a legend on the accuracy. The number can be expressed as a percentage by multiplying the number by 100%. This percentage means that after a certain split, that percentage of data is still kept. So for example, Gini split 5 is 0.968, or 96.8%. This implies that, at the fifth split, the Gini impurity loses 100% – 96.8% = 3.2% of accuracy of the decision tree. Figure 14 shows that the accuracy is taking a horizontal line in accuracy growth after the fourth split for both Gini and Entropy. Within the fourth split, Entropy impurity is better performing with 95.6% in comparison to Gini impurity with 94.1%. Therefore, this research chooses to apply Entropy impurity in splitting the external dataset.

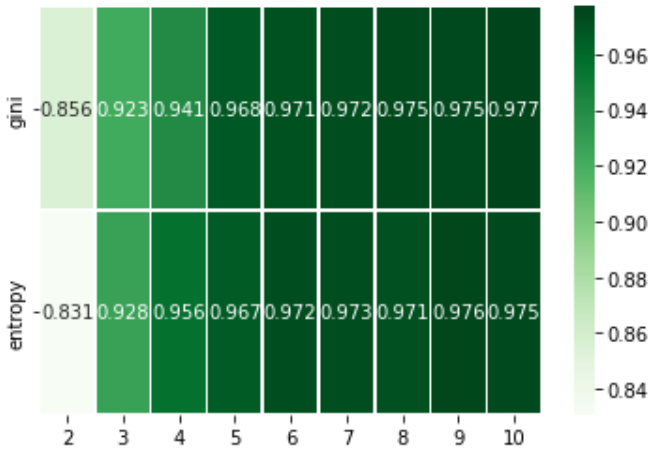


Figure 14 - Heat map showcasing the accuracy for Gini and Entropy for the external dataset

4.1.3 Confusion matrix

A confusion matrix is a performance measurement for machine learning classifications (Shmueli, Bruce, & Patel, 2020). This matrix is used to predict the accuracy of the machine learning model as explained in chapter 3.5.3 in figure 11. Figure 15 shows the confusion matrix based on the external dataset. It shows the True Positive (TP) on the top-left, False Positive (FP) on the top-right, False Negative (FN) on the bottom-left, and True Negative (TN) on the bottom-right.

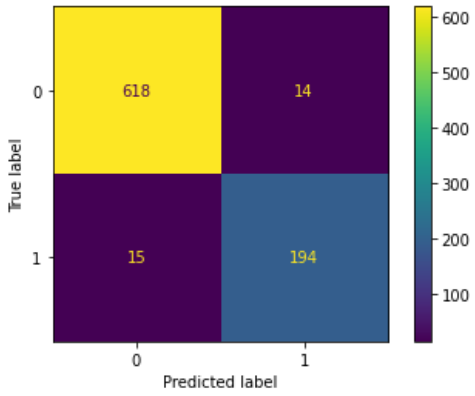


Figure 15 - Confusion matrix for external dataset

Before analyzing the confusion matrix it is important to know that TP and TN sums the correctly predicted classifications, whereas FP and FN represent the falsely predicted classification. This indicates that, the aim of the machine learning model is to maximize the TP and TN, and minimize the FP an FN. Table 1 shows the meaning of every outcome for this model. For example, every True Positive outcome indicates that the model prediction and the actual result is both "left", which means that the employee has left.

Confusion matrix outcomes external dataset

True Positive (TP) = 618	Prediction: employee left | Actual: employee left
False Positive (FP) = 14	Prediction: employee stayed | Actual: employee left
True Negative (TN) = 194	Prediction: employee stayed | Actual: employee stay
False Negative (FN) = 15	Prediction: employee left | Actual: employee stayed

Table 1 - Confusion matrix outcomes for external dataset

Figure 8 shows how to calculate the accuracy according to the confusion matrix. The outcome for accuracy is TP + TN / TP + TN + FP + FN = 618 + 194 / 618 + 194 + 14 + 15 = 96.55%. This indicates that, for 96.55% of the times the model classifies an employee to either "stay" or "leave" the organization, the model will predict it correctly. Whereas for 100% - 96.55% = 3.45% of the times, the model will incorrectly predict whether an employee stays or leaves.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Figure 16 - Formula accuracy based on confusion matrix

4.1.4 Decision tree (external dataset)

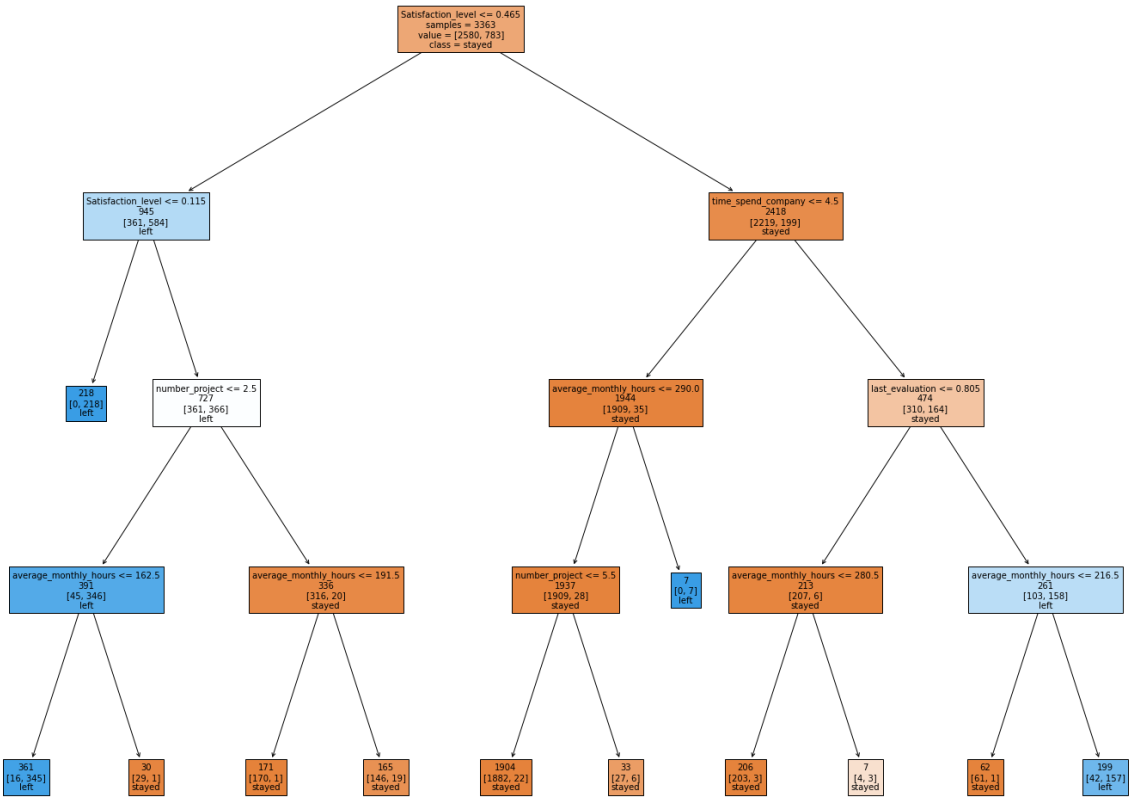


Figure 17 - Decision tree for the external dataset

Figure 17 shows the decision tree based on the external dataset. As decided in chapter 4.1.2, the number of splits is four. This means that there will be a maximum of four yes/no question in order to classify the outcome to either “stayed”, which refers to the that the employee is still working for the organization, or “left”, referring to that the employee has left the organization. There are multiple routes to come to an answer. This depends on the variables that the employee has.

The top-first node is called the root node and is generally considered the most important feature compared with all other features (Shmueli, Bruce, & Patel, 2020). The root node in the decision tree ,based on the external dataset, starts with the satisfaction level, which is confirmed to be the most important feature in Figure 13. The last node is called the leaf node, which classifies the outcome to either “stayed” or “left”. Following is an example of a possible route from root-node to leaf-node:

1st split: satisfaction level <=0.465? is “no”

2nd split: time spend company <= 4.5 is “yes”

3rd split: average monthly hours <= 290.0 is “yes”

4th split: number project <= 5.5 is “no”

Outcome: “stayed”

Another way to write this is as follows:

Satisfaction level > 0.465 AND time spend company <= 4.5 AND average monthly hours <= 290.0 AND number project > 5.5, THEN “stayed”

Application of the decision tree in the future is easy; whenever new data is inserted in the database that is connected to the decision tree, one of the possible routes will be followed from root-node to leaf-node depending on the variables of that particular employee. The outcome will be either “left” or “stayed” and based on that, the user can decide how to react to the situation to change the outcome to its desired outcome.

4.2 Kasparov dataset results

4.2.1 Feature importance

Figure 18 shows a bar chart showing the most important features based on Kasparov’s dataset. It is scaled between 0 and 1, which can be expressed in percentages and therefore, the features are equally divided and will add up to 1, or 100%. The chart shows the important features in descending order and therefore, the top feature is determined to be most impactful in the classification of the outcome.

Compared to the bar chart in figure 13, variables satisfaction level and last evaluation are missing. This is due to Kasparov not collecting these variables from its employees and therefore, are kept out of the set. This implies that, besides these variables, number project, average monthly hours, time spend company, and salary high are the most impactful variables, in descending order. This shows that both datasets are very similar to each other and the features can be generalized in this research. It is clear that both that the findings of feature importance show the same features for both datasets. All other variables show low importance, and therefore will have low impact on the data. This indicates that, only the impactful features will have great impact on classifying outcome for employee.

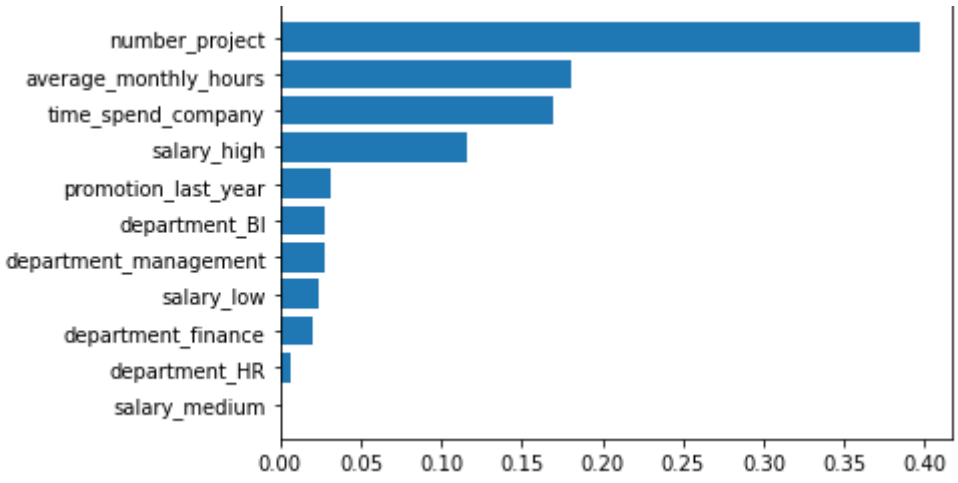


Figure 18 - Bar chart showing the most important features for Kasparov's dataset

4.2.2 Deciding the best split

Like section 4.1.2, determining the best split based on the external set is being replicated on Kasparov’s set. This process is enforced to obtain reliable sub-nodes to deliver an accurate predicting machine learning model (Hssina, Merbouha, Ezzikouri, & Erritali, 2014). The goal is to build a decision tree as pure as possible (Bento, 2021).

The two measurements that are being compared to one another are the Gini impurity and Entropy impurity (Shmueli, Bruce, & Patel, 2020). Its result shows how much data is being kept in the model and how much data is lost in the process of splitting the nodes. Based on those numbers, the number of splits are decided bearing in mind the desire for high accuracy and protecting the model its accuracy.

Figure 19 plots a heat map where the results from both Gini and Entropy are normalized with the same scaler. The x-axis shows the number of splits, the left y-axis shows the method and the right y-axis shows a legend on the accuracy. The number can be expressed as a percentage by multiplying the number by 100%. This percentage means that after a certain split, that percentage of data is still kept.

Figure 19 shows that it takes Gini only three splits to reach > 90% accuracy, whereas Entropy needs four splits to reach that. Comparing the fourth and fifth split for both measurements, Gini increases by 1.2% whereas Entropy decrease 0.2%. The differences are very small and therefore not worth to affect the simplicity of the model. Therefore, the fourth split will be applied, where Gini impurity scores highest with 92.3% in comparison to Entropy with 91.3%. Therefore, this dataset applies Gini impurity in splitting Kasparov's dataset.

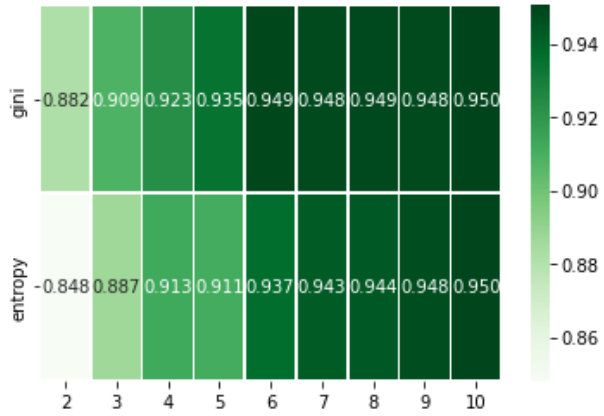


Figure 19 - Heat map showcasing the accuracy for Gini and Entropy for Kasparov's dataset

4.2.3 Confusion matrix

A confusion matrix is used to plot the justification of the accuracy predicted of the machine learning model. Figure 20 shows the confusion matrix for Kasparov's dataset. The numbers in each square represent the number of predictions in either TP, FP, FN, or TN from top-left to bottom-right, respectively. Table 1 shows the meaning of every outcome for this model.

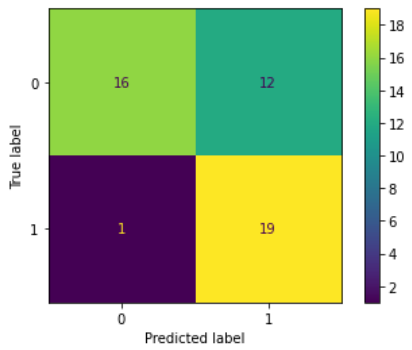


Figure 20 - Confusion matrix for Kasparov's dataset

Figure 8 shows the formula for accuracy based on the confusion matrix. The outcome for accuracy is TP + TN / TP + TN + FP + FN = 16 + 19 / 16 + 19 + 12 + 1 = 81.25%. This indicates that, for 81.25% of the times the model classifies an employee to either “stay” or “leave” the organization, the model will predict it correctly. . Whereas for 100% - 81.25% = 18.75% of the times, the model will incorrectly predict whether an employee stays or leaves.

The outcome in accuracy is significantly different than the accuracy of the external set. The first thing that stands out is the low numbers compared to the external dataset. This is because the significant differences in the number of data points between the two datasets. Chapter 3.2 and 3.3 elaborate on this fact and concludes that the external dataset has 35 times the data that Kasparov has. According to the rule of thumb, a dataset should be at least 10 to 100 times the variables are used to build a good machine learning model, which would be 6 variables times 10 to 100 equals 60 to 600 data points, which is less than the current 48 data points used in the testing set (Alwosheel, van Cranenburgh, & Chorus, 2018).

Figure 20 shows that FP = 12 is significantly higher than FN = 1. This implies that, for 25% of the times the model predicts that an employee stays, but actually will leave (see table 2). This means that the model will not alert the user that an employee is at risk to leave the organization for 25% of the times.

Confusion matrix outcomes Kasparov dataset

True Positive (TP) = 16	Prediction: employee left | Actual: employee left
False Positive (FP) = 12	Prediction: employee stay | Actual: employee left
True Negative (TN) = 19	Prediction: employee stay | Actual: employee stay
False Negative (FN) = 1	Prediction: employee left | Actual: employee stay

Table 2 - Confusion matrix outcomes for Kasparov's dataset

4.2.4 Decision tree (Kasparov's dataset)

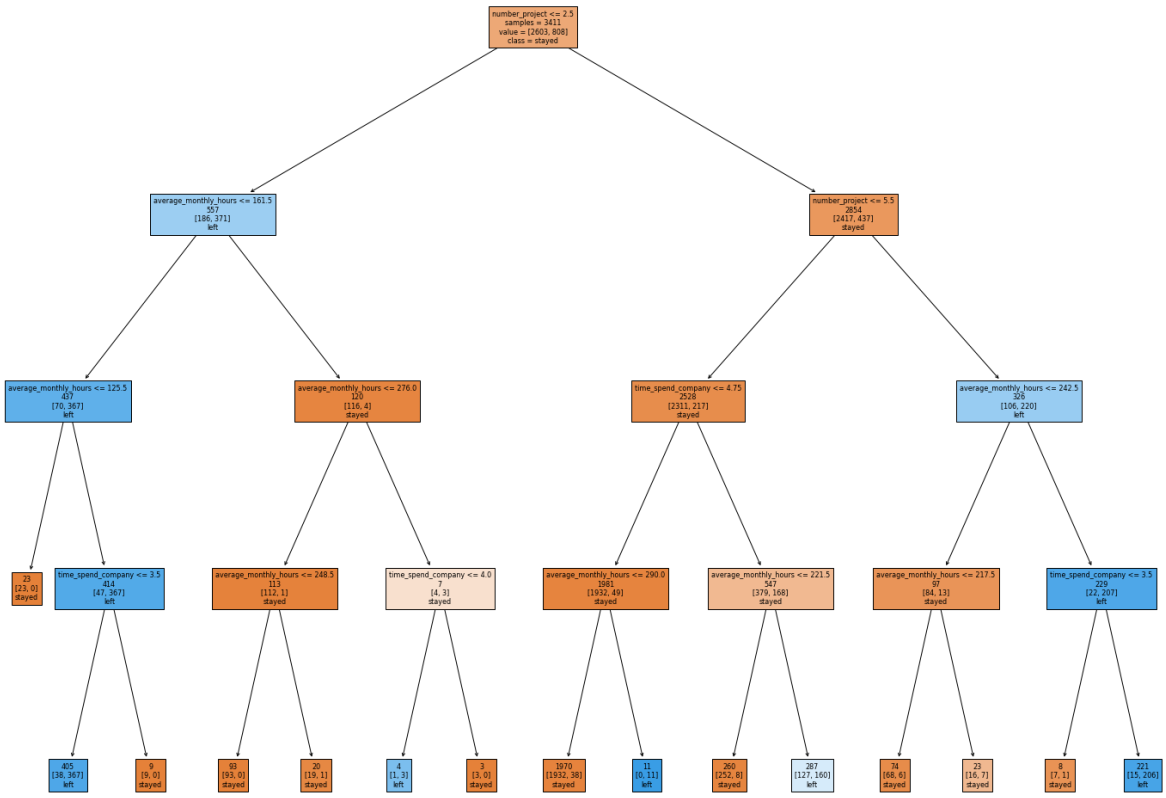


Figure 21 - Decision tree for Kasparov's dataset

Figure 20 plots the decision tree based on Kasparov's dataset. Chapter 4.2.2 concluded that this decision tree will use Gini impurity on four splits. This indicates that, a maximum of four yes/no possibilities will be shown in the decision tree to classify either "left" or "stayed".

Top root node is number project, which is the most important feature compared with all other features in figure 18 (Shmueli, Bruce, & Patel, 2020). Following is an example of a possible route from root-node to leaf-node:

1st split: number project <=2.5 is "yes"

2nd split: average monthly hours <=161.5 is "yes"

3rd split: average monthly hours <=125.5 is "no"

4th split: time spend company <= 3.5 is "no"

Outcome = "stayed".

Another way to write this outcome is as follows:

Number project <=2.5 AND average monthly hours <=161.5 AND average monthly hours >125.5 AND time spend company >3.5 THEN "stayed".

4.3 Conclusion results

This chapter showed all results for the two different datasets being used in this research: (1) the external dataset and (2) Kasparov's dataset. The results are then sub-divided in four steps: (1) mapping the important features, (2) deciding the number of splits, (3) displaying the results of all predictions by the decision tree in a confusion matrix, and (4) plotting and explaining the decision tree.

The most important features are ought to be most impactful on the results on the model's results. Both datasets have the same important features, but the external dataset contains satisfaction level and evaluation time on top of it. This indicates that, the features used on the datasets can be generalized because of the same importance.

Determining the best split is very important because it divides the data into smaller, homogeneous groups to allocate data to one class. However, results in losing data which can decrease the accuracy of the final model. Simplicity and minimizing loss of data is very important to consider during this process. Therefore, Gini impurity and Entropy impurity are compared to calculate the purity of the data bearing in mind the simplicity. For the external dataset, four splits are used based on Entropy impurity with an accuracy of 95.6%. Kasparov's set also used four splits but based on the Gini impurity with 92.3% accuracy.

The results of the final model for the external dataset is 96.55%, which is very well within the expectations in the real-world according to data scientists (Shin, 2020). The distribution of the matrix is very good containing only 29 wrongful classifications divided amongst the False Positive and False Negative. Whereas for Kasparov's dataset the accuracy of the final model is 81.25%, which is below the real-world expectations. This gap can be explained by two reasons: (1) due to a low number of data points provided by Kasparov for this research, and (2) because of missing features satisfaction level and evaluation time in the dataset. The most interesting finding is that for Kasparov, the FP = 12, which indicates that for 25% of the times the model predicts that an employee stays, it actually will leave the organization. This is not desirable for the real-world application of the model since the model's purpose is to alert any employee that is at risk of leaving the organization so the user can act upon that to keep the employee within the firm .

Lastly, the decision trees are plotted and explained so the reader can easily understand how to use it. Whenever new data is inserted in the database that is connected to the decision tree, one of the possible routes will be followed from root-node to leaf-node depending on the variables of that particular employee. The outcome will be either "left" or "stayed" and based on that, the user can decide how to react to the situation to change the outcome to its desired outcome.

Chapter 5

5.Conclusions, limitations, and further research

5.1 Conclusions and limitations

This study examined whether it is possible to use Artificial Intelligence to predict employee turnover in the recruitment industry. A supervised machine learning approach was applied to classify whether an employee will stays within the organization or leaves. Therefore, the main research question during the development of the machine learning model was:

'How can the employee turnover of a recruitment agency be predicted using Artificial Intelligence techniques?'

To answer this research question, the sub-questions have to be answered first. The first sub question, *which Artificial Intelligence techniques can be used to predict outcomes*, can be answered by using theory of chapters 2.1 and 2.2. There are various supervised machine learning techniques but this research chose to test the following algorithms for its accuracy: logistic regression, k nearest neighbors, decision tree, and random forest (Shmueli, Bruce, & Patel, 2020). These algorithms have been selected because it requires a relatively small dataset and because of the familiarity of the techniques by the researcher. The machine learning techniques are all inspired by the book "Data mining for business analytics" by Shmueli (Shmueli, Bruce, & Patel, 2020). These techniques are relatively simple to implement and since the stakeholder Kasparov, needs to continue using this system without hiring an experienced data scientist, this is a good fit.

The second sub question is "*Which variable(s) cover the measure of employee turnover the best?"*. Theory of chapter 2.6 can be used to answer this question. According to Woods, employee turnover measures the ratio between employees that leave and -stay at the organization (Woods, sd). It is used as a tool for management to examine the reasons for the turnover percentage and act accordingly to decrease the outcome. It is desirable that the employee turnover is as low as possible which means that employees are rather staying than leaving. According to various academic journals and -articles from chapter 2.6.3, the following variables have high relevance in employee turnover: salary, secondary benefits, training & development, age & gender, the size of the employer, satisfaction level, promotion, and last evaluation have the most impact on employee turnover.

To answer the research question, this research used two datasets: (1) an external dataset retrieved from the Internet on which the model is built, and (2) Kasparov's dataset which has been used to partially train and -test the model. The external dataset is necessary because Kasparov's data size is too small to build an accurate machine learning model despite the fact that the algorithms require a relatively small dataset (Alwosheel, van Cranenburgh, & Chorus, 2018). The external data scores 96.55% accuracy on the final model, whereas Kasparov scores only 81.25%. This means that according to Shin (2020), the external dataset scored above sufficient accuracy whereas Kasparov scored below sufficient, taking real-world expectations as a standard (>= 90%) (Shin, 2020). Both datasets contain the following six variables and the external dataset contains two more (which are put in brackets), presenting the feature importance from high to low: (satisfaction level), number project, time spend company, (last evaluation), average monthly hours, salary, department, and promotion last year. Kasparov's

dataset misses the first and fourth important feature in the model: satisfaction level and evaluation time. The feature importance show that salary, department, and promotion last year have little impact on the model for both datasets despite the relevance from literature review described in chapter 2.6.3. This means that employee turnover for both the external dataset and Kasparov are not dependent on salary, the department in which they work in, and if they got promoted in the last year.

This research has great insights in predicting employee turnover, however it surely does contain some limitations. Firstly, the small size of the dataset of Kasparov limits this research and its ability to build an accurate machine learning model, which in this case is a decision tree. Therefore, it is recommended that Kasparov retrieves employee data from its existence onwards instead of just the last three years to build a bigger database and therefore, cut out the external dataset. Cutting out the external dataset allows Kasparov to test more of its own collected variables, which are stated in chapter 3.2, instead of complying to the external dataset's variables to respect the similarity of both datasets for generalizability reasons. Secondly, Kasparov scores significantly lower accuracy than the external set. The difference consists mostly out of false positives, which mean that Kasparov's model will not alert users that an employee is at risk of leaving the organization for 25% of the times. This limits the real-world application of Kasparov's model since its purpose is to alert the user when an employee is at risk of leaving so management can act upon it. Thirdly, the external set consists out of two more variables than Kasparov: satisfaction level and last evaluation. These variables are kept in the model because of its high effectiveness on the model. This means that Kasparov is missing out on two variables with high effectiveness on the model's accuracy to predict employee turnover. Based on this, Kasparov should collect this data from its employees in the future to build a better scoring decision tree. As mentioned above, salary, department, and promotion last year have little impact on both datasets. This indicates that this research can conclude that recruitment agencies should focus less on these variables when aiming to prevent employees to leave the organization. Rather, they should focus on the high importance variables: satisfaction level, number project, time spend company, last evaluation, and average monthly hours. This research has tested eight variables for their effectiveness regarding employee turnover. As mentioned above, because this research used two datasets and new variables could not be introduced because this would limit the generalizability of the results, some relevant variables described in chapter 2.6.3 could not be used: secondary benefits, training & development, age & gender, and the size of the employer. This implicates that, this research limits itself to only eight variables where it could be testing and collecting more of these.

5.2 Future research

Summarized, this research provided valuable insights into using machine learning techniques within the recruitment sector. The results from the external dataset proved that with the variables that are used, an accurate decision tree can be built to predict employee turnover. In order to increase the accuracy of Kasparov's dataset, future research should include more data points, each representing an employee. By doing that, not only the second dataset could be excluded, but also more variables could be collected by Kasparov and testing for its relevancy and effectiveness because currently, it is unknown whether these variables have any importance within the model. Lastly, this research has proven that the salary of an employee, the department in which they work in, and whether an employee got promoted within the last year have little

to no effect on the choice of staying or leaving the organization and therefore, can be excluded from future research and lay the focus on other variables that are found important by literature.

References

Ahmad, N., & Daud, S. (2016). Engaging People with Employer Branding. *Procedia Economics and Finance,* Volume 35, pp. 690-698.

Allibhai, E. (2018, October 3). *Hold-out vs. Cross-validation in Machine Learning*. Retrieved from Medium: https://medium.com/@eijaz/holdout-vs-cross-validation-in-machine-learning-7637112d3f8f#:~:text=Cross%2Dvalidation%20is%20usually%20the,just%20one%20train%2Dtest%20split.

Alwosheel, A., van Cranenburgh, S., & Chorus, C. G. (2018). Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis. *Journal of Choice Modelling,* 167-182.

Aly, M. (2005). *Survey on Multiclass Classification Methods.*

Barkved, K. (2022, March 9). *How To Know if Your Machine Learning Model Has Good Performance*. Retrieved from Obviously AI: https://www.obviously.ai/post/machine-learning-model-performance

Bento, C. (2021, June 28). *Decision Tree Classifier explained in real-life: picking a vacation destination*. Retrieved from Towards Data Science: https://towardsdatascience.com/decision-tree-classifier-explained-in-real-life-picking-a-vacation-destination-6226b2b60575

Bento, C. (2021, June 2021). *Decision Tree Classifier explained in real-life: picking a vacation destination*. Retrieved from Towards Data Science: https://towardsdatascience.com/decision-tree-classifier-explained-in-real-life-picking-a-vacation-destination-6226b2b60575

Bini, S. A. (2018). *Artificial intelligence, machine learning, deep learning and cognitive computing: What Do These Terms Mean and How Will They Impact Health Care?* The Journal of Arthroplasty.

BNR Webredactie. (2021, februari 17). *TEKORT AAN FINANCIËLE PROFESSIONALS BLIJFT GROOT*. Retrieved from BNR economie: https://www.bnr.nl/nieuws/economie/10433484/tekort-aan-financiele-professionals-blijft-groot

Bourne, J. (2022, February 4). *What Is Job Satisfaction and Why Is It Important?* Retrieved from Positive psychology: https://positivepsychology.com/job-satisfaction/

Bowen, D. E., & Ledford, G. E. (1991). Hiring for the organization, not the job. *Academy of Management Perspectives,* Vol 5, No. 4.

Branham, L. (2005). *The 7 Hidden Reasons Employees Leave.* New York: AMACOM.

Breiman, L., Friedman, H., Olshen, R. A., & Stone, C. J. (1984, June 3). *lassification and Regression Trees.* CRC. Retrieved from scientist cafe: https://scientistcafe.com/ids/splitting-criteria.html

Brooks, R. A. (1991). *Intelligence without representation.* Cambridge: MIT Artificial Intelligence Laboratory.

Brownlee, J. (2017, July 26). *What is the Difference Between a Parameter and a Hyperparameter?* Retrieved from Machine Learning Mastery: https://machinelearningmastery.com/difference-between-a-parameter-and-a-hyperparameter/#:~:text=Some%20examples%20of%20model%20hyperparameters,k%20in%20k%2Dnearest%20neighbors.

Brownlee, J. (2018, May 23). *A Gentle Introduction to k-fold Cross-Validation*. Retrieved from Machine Learning Mastery: https://machinelearningmastery.com/k-fold-cross-validation/

Buchanan, B. G. (2005). *A (very) brief history of artificial intelligence.* AI Magazine.

Bujang, M., Sa'at, N., Abu Baka Sikdik, M. I., & Joo, L. C. (2018). Sample Size Guidelines for Logistic Regression from Observational Studies with Large Population: Emphasis on the Accuracy Between Statistics and Parameters Based on Real Life Clinical Data. *The Malaysian journal of medical sciences*, 122-130.

Cao, L. (2020). *Data Science: A Comprehensive Overview.* Sydney: University of Technology Sydney.

Clifford, C. (2017, July 17). *Elon Musk: 'Robots will be able to do everything better than us'*. Retrieved from CNBC: https://www.cnbc.com/2017/07/17/elon-musk-robots-will-be-able-to-do-everything-better-than-us.html

Consultancy. (2022, March 16). *Secondary Benefits*. Retrieved from Consultancy: https://www.consultancy.org/career/secondary-benefits#:~:text=Secondary%20benefits%20are%20the%20rewards,or%20the%20infamous%20thirteenth%20month.

Daley, S. (2021, August 9). *28 Examples of Artificial Intelligence Shaking Up Business as Usual*. Retrieved from Built in: https://builtin.com/artificial-intelligence/examples-ai-in-industry

Dawson, S. (2021, October 19). *Importance of Employee Performance Evaluation*. Retrieved from Workast: https://www.workast.com/blog/importance-of-employee-performance-evaluation/#:~:text=Giving%20Clarity%20to%20Employees,their%20sake%20and%20the%20company.

Determinants of Job Satisfaction in the Workplace. (2022, March 16). Retrieved from The Chicago School of Professional Psychology: http://psychology.thechicagoschool.edu/resource/industrial-organizational/determinants-of-job-satisfaction-in-the-workplace

Duin, S. v., & Bakshi, N. (2017, March 28). *The most used terminology around AI*. Retrieved from Deloitte: https://www2.deloitte.com/nl/nl/pages/data-analytics/articles/part-1-artificial-intelligence-defined.html

Dwesini, N. F., & Sisulu, W. (2019). *Causes and prevention of high employee turnover within the hospitality industry: A literature review.* Cape town: African Journal of Hospitality.

Einstein, M. (2019, April 17). *Some Amazing Statistics about Online Data Creation and Growth Rates*. Retrieved from Information Overload Research Group: https://iorgforum.org/case-study/some-amazing-statistics-about-online-data-creation-and-growth-rates/

Ertel, W. (2018). *Introduction to artificial intelligence.* Springer.

European Commission. (2022, April 7). *Internal Market, Industry, Entrepreneurship and SMEs*. Retrieved from European Commission: https://ec.europa.eu/growth/smes/sme-definition_en

Fortune Business Insights. (2021, September). *Artificial Intelligence Market*. Retrieved from Fortune Business Insights: https://www.fortunebusinessinsights.com/industry-reports/artificial-intelligence-market-100114

Gebregziabher, D., Berhanie, E., Berihu, H., Belstie, A., & Teklay, G. (2020). The relationship between job satisfaction and turnover intention among nurses in Axum comprehensive and specialized hospital Tigray, Ethiopia. *BMC Nursing*, 79.

GitHub. (2021, February 9). *Employee turnover prediction*. Retrieved from GitHub: https://github.com/Mario-apk/Employee-Turnover-Prediction

Gordon, J. S. (2020). *Smart Technologies and Fundamental Rights.* Brill.

Gregory, K. (2011). *The Importance of Employee Satisfaction*. Retrieved from Neumann University: https://www.neumann.edu/academics/divisions/business/journal/Review2011/Gregory.pdf

Gupta, S. (2019, May 29). *KNN Machine Learning Algorithm Explained*. Retrieved from Springboard blog: https://in.springboard.com/blog/knn-machine-learning-algorithm-explained/

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques.* Elsevier: Amsterdam.

Haoyong, L., & Hengyao, T. (2011). *Machine learning methods and their application research.* Huanggang: International Symposium on Intelligence Information Processing and Trusted Computing.

Hickey, S. (2020). *An investigation into employee turnover in the recruitment industry.* National college of Ireland.

Hollman, K. W., & Abassi, S. M. (2000). Turnover: The real bottom line. *Public Personnel Management*, 333-342.

Hssina, B., Merbouha, A., Ezzikouri, H., & Erritali, M. (2014). A Comparative Study of Decision Tree Id3 and C4.5. *International Journal of Advanced Computer Science and Applications(IJACSA)*.

Idson, T. L. (1993). *Employer size and labor turnover.* Columbia: Columbia University.

India Filings. (2022). *Salary Income under Income Tax*. Retrieved from India Filings: https://www.indiafilings.com/learn/salary-income-under-income-tax/

Koyejo, O., Ravikumar, P., Natarajan, N., & Dhillon, I. S. (2014). *Consistent Binary Classification with Generalized Performance Metrics.*

Kraiger, K., & Aguinis, H. (2009). Benefits of training and development for individuals and teams, organizations, and society. *Annual Review of Psychology*, Vol. 60 pp. 451-474.

Krajcsák, Z., & Kozák, A. (2018). The impact of labor shortage on the employee commitment. *Journal of Human Behavior in the Social Environment*, 1060-1067.

Kuehner-Hebert, K. (2019, April 17). *Promotions play a key role in employee turnover*. Retrieved from BenefitsPro: https://www.benefitspro.com/2019/04/17/promotions-play-a-key-role-in-employee-turnover/?slreturn=20220413050620

Kumar, N. (2019, January 25). *KNN Algorithm in Machine Learning*. Retrieved from The Professionals Point: http://theprofessionalspoint.blogspot.com/2019/01/knn-algorithm-in-machine-learning.html

Lee, S. Y., & Whitford, A. B. (2008). Exit, voice, loyalty and pay: Evidence from the public. *Journal of Public Administration Research and Theory*, 647-671.

Meijer, E. (2022, January 25). *'Krapte op IT-arbeidsmarkt neemt in 2022 verder toe*. Retrieved from AG Connect - The amazing world of IT: https://www.agconnect.nl/artikel/krapte-op-it-arbeidsmarkt-neemt-2022-verder-toe

Narkhede, S. (2018, May 9). *Understanding Confusion Matrix*. Retrieved from Towards Data Science: https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

Peffers, K. (2006). *THE DESIGN SCIENCE RESEARCH PROCESS: A MODEL FOR PRODUCING AND PRESENTING INFORMATION SYSTEMS RESEARCH.* Claremont, CA: DESRIST.

Pratt, K. S., & Murphy, R. R. (2012). Protection from Human Error: Guarded Motion Methodologies for Mobile Robots. *IEEE Robotics & Automation Magazine*, 36-47.

Price, J. L. (1977). *The study of turnover.* Iowa: Iowa State University.

PwC. (n.d.). *PwC's Global Artificial Intelligence Study: Exploiting the AI Revolution*. Retrieved from PwC: https://www.pwc.com/gx/en/issues/data-and-analytics/publications/artificial-intelligence-study.html

Python. (2021, October 21). *Python 3.10.0*. Retrieved from Python: https://www.python.org/downloads/release/python-3100/

Rimol, M. (2021, November 22). *Gartner Forecasts Worldwide Artificial Intelligence Software Market to Reach $62 Billion in 2022*. Retrieved from Gartner: https://www.gartner.com/en/newsroom/press-releases/2021-11-22-gartner-forecasts-worldwide-artificial-intelligence-software-market-to-reach-62-billion-in-2022

Rothwell, W. (2016). *Effective Succession Planning: Ensuring Leadership Continuity and Building Talent from Within.* New York: AMACOM.

Sarker, I. H. (2020). *Mobile Data Science and Intelligent Apps.* Mobile Netw Appl.

Sarker, I. H. (2021). *Deep Cybersecurity: A Comprehensive Overview from Neural Network and Deep Learning Perspective.* SN Computer Science.

SAS. (2022, March 25). *A guide to machine learning algorithms and their applications*. Retrieved from SAS: https://www.sas.com/tr_tr/insights/articles/analytics/machine-learning-algorithms-guide.html

Shin, T. (2020, September 2). *How I Consistently Improve My Machine Learning Models From 80% to Over 90% Accuracy*. Retrieved from Towards Data Science: https://towardsdatascience.com/how-i-consistently-improve-my-machine-learning-models-from-80-to-over-90-accuracy-6097063e1c9a

Shmueli, G., Bruce, P. C., & Patel, N. R. (2020). *Data mining for business analytics.* Frankfurk at Main: Wiley.

Staffing Industry. (2022, January 25). *TEMPORARY EMPLOYMENT SECTOR REPORTS GROWTH IN 2021 WITH HOURS AND TURNOVER ON THE RISE*. Retrieved from Staffing Industry: https://www2.staffingindustry.com/eng/Editorial/Daily-News/Netherlands-Temporary-employment-sector-reports-growth-in-2021-with-hours-and-turnover-on-the-rise-60396

Taylor, S. (2014). *Resourcing and Talent Management.* th ed. London: Chartered Institute of Personnel and Development.

TechTarget. (2022, April 28). *Validation set*. Retrieved from TechTarget: https://www.techtarget.com/whatis/definition/validation-set

The History of Artificial Intelligence. (2017, August 28). Retrieved from Harvard University: https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/

van Vuuren, P. (2022, May 17). *Krapte arbeidsmarkt neemt toe, aantal banen op recordniveau*. Retrieved from Accountant: https://www.accountant.nl/nieuws/2022/5/krapte-arbeidsmarkt-neemt-toe-aantal-banen-op-recordniveau/

Vardi, Y. (1980). Organizational career mobility: An integrative model. *Academy of management review*, vol 5. 341-356.

Venable, J., & Baskerville, R. (2012). Eating our own Cooking: Toward a More Rigorous Design Science of Research Methods. *Electronic Journal of Business Research Methods*, 141-153.

Victoria Séveno. (2021, November 16). *The Dutch labour crisis continues: 126 jobs per 100 unemployed*. Retrieved from I am Expat: The Dutch labour crisis continues: 126 jobs per 100 unemployed

Wakefield, K. (2022, March 25). *A guide to the types of machine learning algorithms and their applications*. Retrieved from SAS:

https://www.sas.com/en_gb/insights/articles/analytics/machine-learning-algorithms.html

Werbel, J. D., & Bedeian, A. G. (1989). Intended turnover as a function of age and job performance. *Journal of organizational behavior*, VOL. 10, 275-281.

Woods, C. (n.d.). *What Is Employee Turnover? - Definition, Cost & Reasons*. Retrieved from Study: https://study.com/academy/lesson/what-is-employee-turnover-definition-cost-reasons.html

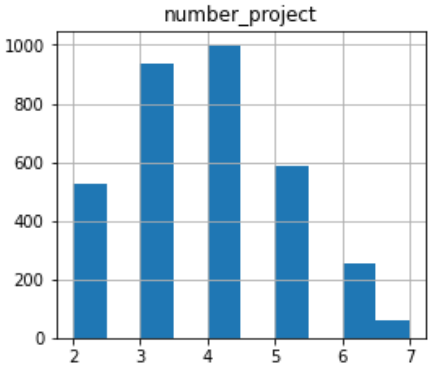
Yildirmaz, A., Ryan, C., & Nezaj, J. (2019). *State of the Workforce Report.* ADP Research Institue.

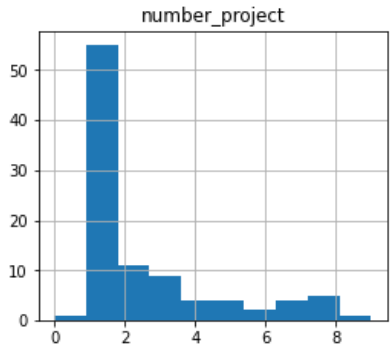
Yiu, T. (2019, June 12). *Understanding Random Forest*. Retrieved from Towards Data Science: https://towardsdatascience.com/understanding-random-forest-58381e0602d2#:~:text=The%20random%20forest%20is%20a,that%20of%20any%20individual%20tree.

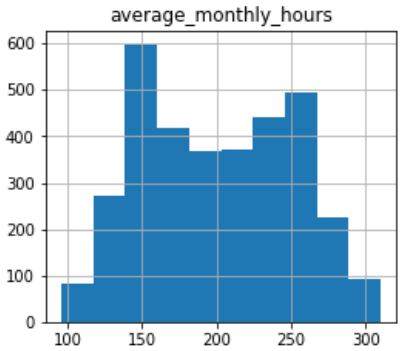
Zakariah, M. (2014). Classification of large datasets using Random Forest Algorithm in various applications: Survey. *International Journal of Engineering and Innovative Technology (IJEIT)*, Volume 4, Issue 3.

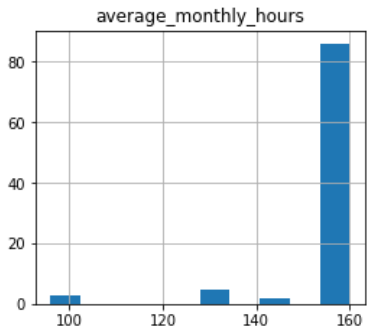
Appendices

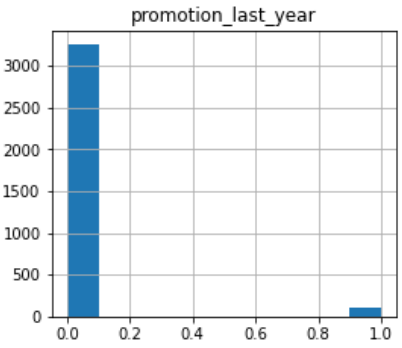
Appendix A

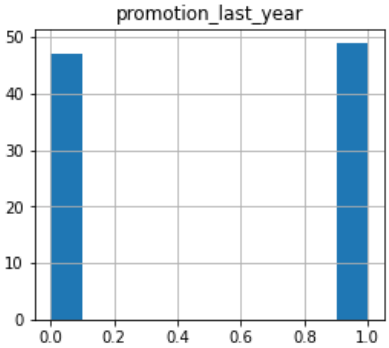


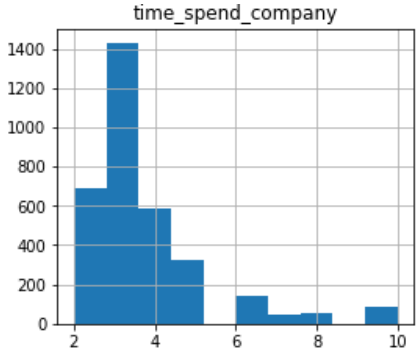


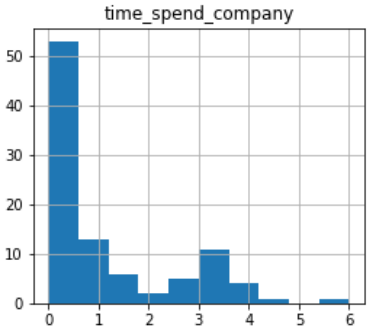


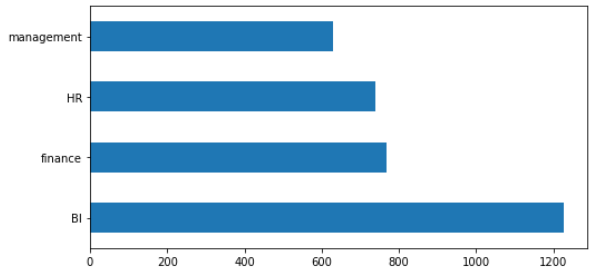


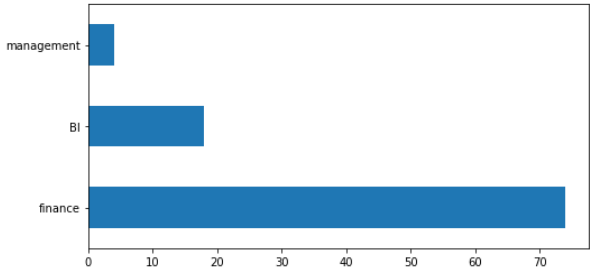


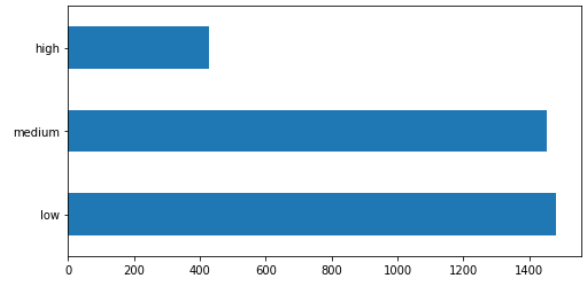


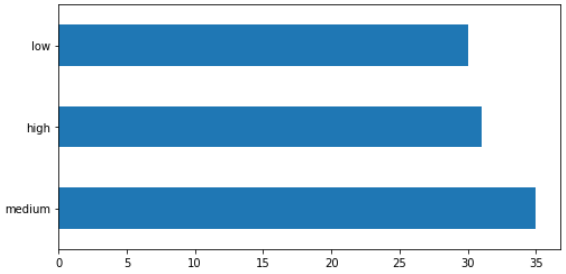





management
HR
finance
BI
0
200
400
600
800
1000
1200


management
BI
finance
0
10
20
30
40
50
60
70


high
medium
low
0
200
400
600
800
1000
1200
1400


low
high
medium
0
5
10
15
20
25
30
35