NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

Identifying and characterizing employee groups by turnover risk using predictive analytics

Bruno Cassanta Vidotto

Project Work presented as partial requirement for obtaining the Master's degree in Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

IDENTIFYING AND CHARACTERIZING EMPLOYEE GROUPS BY TURNOVER RISK USING PREDICTIVE ANALYTICS

Case study of a European Multinational Company

by

Bruno Cassanta Vidotto

Project Work presented as partial requirement for obtaining the Master's degree in Information Management, specialization in Knowledge Management and Business Intelligence

Advisor: Prof. Mijail Juanovich Naranjo-Zolotov

February 2021

ACKNOWLEDGEMENTS

Thanks to my family who has always been supportive of my studies and believed in me: my father Juarez, my mother Sandra, and my sister Laís. Also, a special acknowledgment and thank you to my late grandfather Josué Vidotto, who had to make hard choices in his life to ensure the education of his family during difficult times and has always been an inspiration.

Thanks to my girlfriend Giovanna, who was by my side in the time of the writing, throughout these challenging months.

Thanks to the leaders of the studied organization, who were open, allowed this study and the exploration of this new approach and mindset.

And, lastly, a special thanks to my advisor Prof. Mijail, who from the beginning, supported and guided me through this journey and to all of my professors in NOVA IMS for the knowledge shared during the master program.

ABSTRACT

This project presents a predictive analytics project developed in a European multinational to understand and predict the turnover of its employees. It analyses the Human Resources current challenges, such as the increasing global competition for talent, where players compete for scarce skillsets such as technology and data science, and the new strategies necessary to deal with this scenario. The study explores the literature review of these contextual matters and of the studies of variables that influence turnover, generating insights and input for applying techniques aligned with the new mindset of identifying 'flight-risk' groups and developing targeted actions instead of only one-size-fits-all solutions. The project gathered data from different sources of the organization, designed variables, based on a literature review and internal brainstorms, treated data quality issues, transformed the data and applied three different machine learning algorithms to develop a classification predictive model. The study evaluated 46 input variables and selected a set of 26 that had higher impact on the turnover which were used in the models. Finally, it applied clustering techniques to divide employees in clusters, and identified two containing more extreme turnover behaviors ("Loyal" and "Flight risk") and described them accordingly to their main characteristics contributing with practical insights to support potential decisions.

KEYWORDS

Turnover; Attrition; Human Resources; People Analytics; Machine Learning; Predictive Analytics.

INDEX

1. Introduction 7
   1.1. Context 7
   1.2. Problem justification 8
   1.3. Research objectives 8
   1.4. Introduction to the methodology 9
2. Literature review 10
   2.1. The talent market landscape 10
   2.2. New human resources management mindset 10
   2.3. Understanding and influencing turnover 11
   2.4. Factors influencing turnover 13
3. Methodology 16
   3.1. Frame the problem 17
   3.2. Get the data 17
   3.3. Explore the data 18
   3.4. Prepare the data 20
      3.4.1. Feature engineering 20
      3.4.2. Data cleaning 22
      3.4.3. Feature selection 22
      3.4.4. Data partitioning 24
   3.5. Model the data and fine-tune the models 24
      3.5.1. Random Forest 25
      3.5.2. Logistic Regression 27
      3.5.3. Neural Network 28
      3.5.4. Models results 29
   3.6. Present the solution 30
   3.7. Launch of the ml system 33
4. Conclusions 34
5. Limitations and recommendations for future works 35
6. Bibliography 36
7. Appendix 39

LIST OF FIGURES

Figure 1 - Machine learning project steps 16
Figure 2 - Histogram of variable age 18
Figure 3 - Age of individual contributors and leaders 18
Figure 4 - Turnover by gender 19
Figure 5 - Proportion of turnover by management level 19
Figure 6 - Turnover by tenure (time in company) by time in the same job profile 20
Figure 7 - Correlation matrix 23
Figure 8 - “Ideal exercise” decision tree example - created by the author 25
Figure 9 - Variables importance according to random forest algorithm 26
Figure 10 - Logistic function (Gerón, 2017) 27
Figure 11 - Perceptron, the simplest artificial neural network structure (Gerón, 2017) 28
Figure 12 - Confusion matrix quadrants, created by the author 29
Figure 13 - Optimal number of clusters 31

LIST OF TABLES

Table 1 - Turnover variable groups 14
Table 2 - Input variables used 21
Table 3 - Data partitioning 24
Table 4 - Most impactful variables according to Logistic Regression 28
Table 5 - Model evaluation 30
Table 6 - Cluster sizes 31
Table 7 - Turnover by cluster 31
Table 8 - Cluster comparison 32
Table 9 - Literature review of variables impacting turnover 39

1. INTRODUCTION

1.1. CONTEXT

The current context presents significant challenges to every organization in terms of people related topics, from attracting to retaining the right talent, going through the entire employee journey. Already in 1998, Chambers *et al.,* a group of researchers from McKinsey, researched and warned about the reality of a 'war' for talent. They unveiled that the complexity of the skillsets of the talents required was increasing, with the talent supply remaining limited, and that the globalization was making the demand fiercer, where companies from the entire world were increasingly competing for the same talent pool.

This reality, with elements of new skillsets and globalization continued to evolve, making talent competition a common topic within the Human Resources agenda. To deal with it a shift in mindset was and still is necessary. Initially, the standard was holding huge retention programs to keep everyone in the company reducing overall turnover. Now, though, focus should be narrowed, aiming to influence who leaves and when. In summary, it is shifting from a mindset of tending to a water dam and keeping a full water reservoir, to managing an ever-flowing river (Cappelli, 2000).

The tools to adopt this new mindset are available now. Baek (2016) explains that people analytics provides HR with the conditions to plan the necessary targeted actions using a scientific and evidence-based decision-making approach, instead of only intuition and 'gut feel'. And there are tangible results from it, a Deloitte study in 2014 identified that the 14% of companies with mature predictive and strategic talent analytics capabilities outperformed by 30% the S&P 500 from 2013 to 2016. In the many different dimensions of human resources management (Recruitment, Learning and Development, Compensation, Health and Wellness, Leadership, etc.), the needs are evolving from basic descriptive questions into asks for targeted, data-oriented actionable insights, which provide more precise recommendations to help executives and managers to run their business (Guenole *et al.,* 2017).

Large firms like Google, WL Gore, Tesla, and many others now use big data, predictive analytics, and machine learning techniques to monitor and analyze their talent. This gives them the ability to make better decisions on all processes, including how to recruit, onboard, retain, develop, and motivate their people (Schweyer, 2018). According to IBM Executives, for example, they have developed a model that can predict with a 95% accuracy who will quit (Rosenbaum, 2019).

This is a context that demands a lot from business leaders and Human Resources professionals. Understanding and being able to apply the new technologies will be a matter of survival for the organizations in the coming years. Because of this, a structured approach for any analytics effort is necessary in order to avoid pitfalls, such as: starting from the data and not the business questions, lacking or having weak hypothesis, unengaged stakeholders, inaccessible or bad data. All the mentioned factors might lead to the waste of precious resources from organizations (Jain, 2015).

1.2. PROBLEM JUSTIFICATION

Business and Human Resources leaders need to address the challenging scenario presented. Increasing competition for rare talent with specialized skillsets is a reality and the costs of replacing talent are significant. Studies show that each departure costs approximately one third of the workers annual earnings on average, considering hard costs such as temporary worker replacements and recruiting, and soft costs as reduced productivity and time spent interviewing (Agovino, 2019). People analytics, with advanced techniques, such as machine learning, emerges as a tool to support businesses to understand and deal with these issues, supporting retention efforts, for example. The studied organization seats within this environment. It is a European multinational company, inserted in a fast paced, innovative market, in very quick expansion both in terms of sales and employees, creating and structuring its processes, defining and constantly changing its organizational structures and systems, with constant large projects such as acquiring new companies. It is competing for highly demanded talents, such as tech and data specialized professionals who are scarce everywhere in the world, and the "fight" for these skillsets is against giants such as Google and Facebook. This case study utilizes predictive analytics techniques and develops its application in a complex organization to create a data-oriented basis for answering the following problem: Who are the employees that have the highest and lowest turnover probability in the studied organization?

1.3. RESEARCH OBJECTIVES

The context, the justification and problem set the needs to be tackled within this study. Considering key factors such as the challenging competitive talent market, the evolving practices of Human Resources management, the artificial intelligence discoveries and tools and the specific environment of the studied company this study aimed at achieving the following objectives:

General Objective: Identify and characterize employee groups with the highest and lowest turnover risks using predictive analytics in a European Multinational.

Specific Objectives:

- Create an employee dataset using data from different sources and references in literature
- Identify the most relevant variables that influence employee turnover
- Build a turnover predictive model testing different algorithms
- Divide employees in clusters and characterize them by turnover risks

1.4. INTRODUCTION TO THE METHODOLOGY

The methodology adopted was an eight-steps approach for machine learning projects, created for making sure the efforts deliver the most value stakeholders (Gerón, 2017).

It started with better framing and understanding the problem, from a business perspective, based on the context, justification and objectives mentioned and using insights from the literature review.

Afterwards, data related steps included: obtaining the data and structuring the dataset from the sources available, exploring it and preparing it for the modelling. Here, pre-processing procedures to improve data quality (e.g. removing outliers, dealing with missing values) and data transformation procedures (e.g. transforming, creating variables) were conducted, based on the literature and inputs from stakeholders within the organization. Here, also, a dimensionality reduction process was done to select the most relevant dimensions impacting the target variable.

Then, there were the modeling steps when predictive algorithms were used to model the data and fine-tune the best performing modules. In these steps, different predictive machine learning supervised algorithms were tested to predict turnover, using as input the dataset with the previously prioritized variables.

The last set of activities were to conclude the analysis and present the solutions to the problem. A non-supervised machine learning clustering technique was used to group the observations in clusters, which were later described and characterized, according to the prioritized variables and their overall turnover behaviors.

In the end, recommendations, and limitations for applications in “business as usual” situations were given, and the conclusions were organized. The last step in the methodology, which is launching the machine learning system was out of the scope of this study, but suggestions were given to support its execution.

2. LITERATURE REVIEW

2.1. THE TALENT MARKET LANDSCAPE

The current talent market is increasingly challenging from two angles: globalization makes competition become fiercer and speed of changes make the 'content' of the talents themselves change faster. Chambers *et al.,* mention that more sophisticated talents are being demanded such as multi-cultural experience, technological and entrepreneurial skills, ability to thrive in unstructured organizations and the concrete specific needs change in a very fast pace.

In more recent years, almost no highly skilled worker is involuntary without work, causing the need for an extra effort from organizations to retain current talent, especially in terms of work design and work environments (Schweyer, 2018). Employees with high-demand or difficult-to-replace-skills (e.g. computer scientists) or high performers are not impacted by tighter labor markets, even in high unemployment rates contexts, so their turnover costs are higher (Allen *et al*., 2017).

Besides the evolution of the knowledge and skills needs, the rise of many small and medium-sized companies represents new entrants in the competition for talent. They bring a different value proposition to employees and increase turnover risks (Chambers *et al.,* 1998*)*.

Nowadays, every company is becoming a talent poacher, looking outside the organization to spot great candidates and lure them from their current employers. This scenario is not limited geographically or by sector since they can be linked to specific functions or skillsets. For example, an executive from Marriot hotels with extensive experience in customer service could be suited to lead a service improvement effort in an airline (Cappelli, 2000). Also, in general, people are more prone to job mobility. While, in the past, a person with a high performance might have changed once or twice from jobs in their lifetime, now that number is likely to be much higher (Chambers *et al.,*1998*)*.

The factors presented add up to a challenging scenario. Talent is more complex and sophisticated, therefore becomes scarcer, at the same time, globalization and function specialization increase the number of potential competitors fighting to attract and keep the talent. Now, there is a scenario where traditional companies in the US must compete with start-ups in Australia, or Germany, for example. All of that in a scenario where company loyalty by employees is reducing.

The employee retention topic, within this context, is critical for organizations and their leaders. The costs of replacing an employee, including recruiting, selecting, and training new employees, often exceed 100% of the annual salary for the position being filled, besides the potential loss of knowledge, client satisfaction and other factors that cannot be measured (Cascio, 2006) . That leads to a need for a new human resources management mindset.

2.2. NEW HUMAN RESOURCES MANAGEMENT MINDSET

Cappelli (2000) explains that the main assumption now is that long-term, flat employee loyalty is neither possible nor desirable. The author uses a metaphor of a reservoir damn to represent the old *status quo*, where the focus in keeping the water in. Now the new mindset is one of an ever-flowing river, in which management aims to control its direction and its speed. In other words, instead of

having broad retention programs that reach every employee to keep them all, the effort should be turned into highly targeted efforts focused on particular employees or groups. Now human resources management aims to influence who leaves and when.

The old paradigm was under the assumption that a simple one-size-fits-all was the most effective retention strategy. However, the new evidence-based approach suggests the opposite, arguing that actions require a deeper diagnostic, to identify the kinds of turnover, if they are a problem and what are their specific causes. This continuous process of understanding the broad context, both internal and external, and deep diving using data to identify and validate underlying causes, is crucial to generate input for designing effective targeted and organized retention initiatives (Allen *et al*., 2017). Schweyer (2018) develops the argument even further mentioning that, in the current context, organizations should understand the unique motivators of each employee, as in the trend with customers management, and create highly engaging, hyper-personalized work experiences.

Organizations that define, develop, and deliver the best employee value proposition will be able retain people the desired people (Chambers *et al.*, 1998). As in a value proposition to a customer, it involves understanding specific needs of that segment, group of people, and setting your processes to deliver accordingly, it is a matter of having a clear focus.

In the beginning of the twentieth century, Taylor's and Ford's experiments starting "scientific management" measured people related metrics of performances to optimize results. These initial measurements were input to job design, however, then, the focus was merely reducing time for operational tasks and outputs, resulting sometimes in the alienation of workers. Now, with the given context of evolving human resources practices and technological changes, innumerous possibilities of new quantitative analysis exist, but. this time, potentially focusing on other many dimensions of employee experience (e.g. well-being, diversity, development). Nevertheless, new ethical challenges come with this exponential increase of people's data points, potentially affecting people's privacy and, therefore, trust, and need to be taken into consideration (Chamorro-Premuzic and Bailie, 2020).

Given this potential of a more data-driven human resources approach, Morrison (2018) raises another argument, of the need for Human Resources to have a more strategic role. He argues HR needs to use the data-driven insights to partner with the business and take a more proactive role shaping the strategy. This includes understanding the changing market and employee needs and planning ahead, evaluating alternative future scenarios of workforce needs and strategies in term of size, organizational design and skills, and support more concrete, data-driven decisions, adopting a role similar to the role of Finance Planning & Analytics (FP&A) teams play in Finance broader departments.

2.3. UNDERSTANDING AND INFLUENCING TURNOVER

The turnover topic, also called attrition, is part of this new strategic data-driven HR agenda. The German company SAP people analytics team is an example of this and used people analytics to analyze the turnover problem. It generated insights using clustering techniques and was able to partner with the business for more solid structured decision making, linking solutions with broader business and financial impacts, such as ROI estimations (Becker, 2019). To achieve similar results in a

consistent way, it is necessary to understand factors behind turnover in order to be able to influence it.

The employee turnover phenomenon is studied since the beginning of the century, with its first empirical study being published in 1925 involving clergyman and evolved through time in several different applications and environments (Hauskneckt *et al.*, 2017). During the investigation of turnover in an organization a premise must be clear: not all turnover is the same. An initial question to frame the issue is 'who is leaving?'. Are individuals of highly pivotal positions? Are they high performers? If yes to the two latter questions, it surely is a problem, but if not, it may be good for the organization. In the diagnosis is necessary to also consider the turnover costs, evaluating what the organization is losing and what investment (e.g. time, money) would be necessary to design an action to reduce turnover (Allen *et al*., 2017).

Rao (2007) suggests initial exploratory questions to identify the focus of turnover. If it is focused in some discipline, function, department, section, region, linguistic group, gender. If it can be sliced by reasons, by controllable or non-controllable, by colleges, institutions, or age groups. Applying this kind of intelligence, slicing data, to understand it and describe it represents an application of People Analytics, which is the opportunity for HR practitioners and managers to shift from intuition and "gut feel" towards a data-driven approach for people-related decisions (Baek, 2016).

Organizations have the misconception that the only reason why people leave is pay. Payment matters, but topics like job design, growth opportunities and relationship with direct manager are strongly linked to retention or turnover. Decisions to leave, also are often initiated by a shock, an event that triggers the quitting thought by the employee (Allen *et al*., 2017).

Many different factors can influence that decision: individual related, role or job related, organization related, professional, societal factors including peer pressure factors and socio-economic environment related factors (Rao, 2007). Strategies like thinking carefully about which tasks to include in which jobs, or how to increase a sense of belonging and boost colleagues' relationships (social ties) are actions that attack other potential causes of attrition, many times ignored (Cappelli, 2000). Rao suggests for organizations to focus on what they can change and clearly segment the workforce into: those whom the organization wants to have indefinitely, those who should be there for a shorter time and those who should not receive retention efforts at all.

Allen *et al.* describe two different groups of retention actions: systemic, which result in general principles and rules for the broad organization; targeted, which are based on specific drivers and often influence turnover among a specific group of employees. The decision of which kind of action to take depends on the root causes identified and a cost benefit analysis. For example, a systemic action could be changing the compensation structure, if the organization sees that this cause is generating turnover broadly, or as a targeted action, it could be changing the leader of a specific department that is the cause of a localized problem of high turnover.

Independently of the context, the first step to deal with a turnover problem is to gather data and kickoff the people analytics effort. A good starting point is with simple spreadsheets, structured and relevant questions, then good communication and change management (Baek, 2016). Allen *et al.* list as potential data sources exit interviews, post-exit surveys, focus groups, linkage research (gathering data from different sources). Baek identifies as latest trend the concept of the quantified employees

were data comes from them in real time, from internal systems or from employees' external gadgets. But those latest sources, like social media, come with an increasing concern with data confidentiality.

Firms like Google, WL Gore, Tesla, and dozens of others are now using their data to apply predictive analytics, using machine learning techniques to analyze their talent constantly and pro-actively make better decisions in people processes. These leaders are eager to go beyond traditional HR analysis that look at the past (e.g. descriptive). Employees give many signals about their intentions and, with the combination of the increasing amount of data available measuring employees features, it is possible to build predictive models to understand and forecast turnover. HR and all leader can then act developing tailored actions to intervene (Schweyer, 2018). Image 1 identifies the different stages of people analytics ranging from traditional analytics, where the most companies are, to advanced analytics which involve the mentioned predictive and even prescriptive analytics.

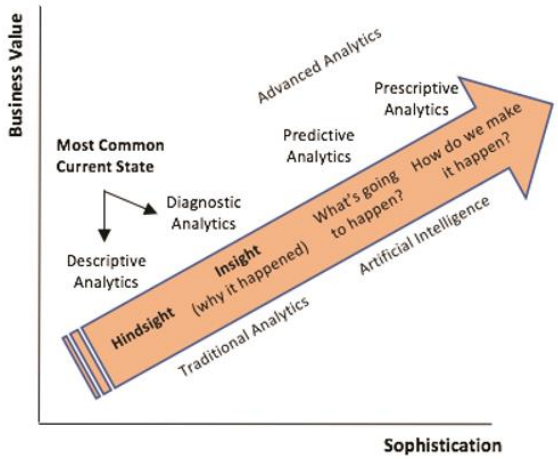

Image 1 - The State of People Analytics (Schweyer, 2018)

Similar steps are identified in a research by Mckinsey (Ledet *et al*, 2020), evolving from turning poor data into high quality data, moving towards performing advanced analytics and reliable predictions with supervised and unsupervised machine learning techniques. The article explains that the journey is iterative, sometimes tackling an issue in an early stage (e.g. data quality foundations), but always moving forward focusing on generating actionable insights for decision makers, being a partner from the business and generating value to the people and the organization.

2.4. FACTORS INFLUENCING TURNOVER

Variables that affect turnover have been studied for a long Time. In 1969, Farris made an extensive work with 395 scientists in two organizations, using surveys proposing and testing statistically hypothesis. Kirschenbaum, and Weisberg in 1990 performed a test that evaluated Israel's textile industry, however, instead of using only intentions they followed up and included the actual separations. This turnover related discussion has also become part of the business discussions as mentioned by Chambers *et al,* in the Mckinsey article in 1998 which utilized varied methods such as qualitative surveys with top level executives, focus groups, interviews with HR executives. In 2015, Rubenstain *et al.* developed a meta-analysis study focused on antecedents of voluntary turnover.

More recently, the artificial intelligence approached were included in the studies, applying machine learning techniques such as the study from Fan *et al.* in 2012, which already included the predictive element using surveys of Taiwan Tech companies. In a similar manner, using partially real data and partially a dataset created by Watson, the Artificial Intelligence from IBM, Zhao *et al.* (2018) also applied advanced analytics from turnover prediction purposes.

As an evolution for the knowledge in human resources management practices and artificial intelligence, this study aims to apply machine learning techniques to solve a real problem using a real dataset of a company with a high level of complexity, with presence in many countries and with a diverse workforce.

Below, Table 1 presents a summary of variable groups that affect the turnover phenomena. The table, created by the author during the literature review, identifies the authors of the studies that mention variables which sit within those variable groups. The 14 groups were created to facilitate the understanding. The full list is available in Table 9 in the appendix. This literature review of variables was used as an input for the creation of the actual variables in the predictive models, later complemented by discussions within the studied organization.

Table 1 - Turnover variable groups

Variable Group	Description	Authors
Work Environment	Variables related with physical work environment and people's perception about the work environment	Ajit, P. (2016) Baek, P. (2016) Cappelli, P., (2000)
Withdrawal Process	Job search behaviors and surveys about people intentions to leave	Allen, D. G., Bryant, P. C. and Vardaman, J. M. (2017) Bretz, R. D., Boudreau, J. W., Judge, T.A. (1994) Kirschenbaum, A., Weisberg, J. (1990) Schweyer, A. (2018)
Team Factors	Variables related to team characteristics and perceptions of employees about their teams	Allen, D. G., Bryant, P. C. and Vardaman, J. M. (2017) Baek, P. (2016) Cappelli, P., (2000) Chambers, E., G., Foulon, M., Handfield-Jones, H., Hankin, S. M., Michaels III, E.G. (1998) Elvira and Cohen (2001) Farris, G. F. (1969) Kirschenbaum, A., Weisberg, J. (1990) Rao, T.V. (2007) Zhao, Y., Hryniewicki, M. K., Cheng, F., Fu, B., & Zhu, X. (2018)
Stability	Employees perception about job stability in the company.	Cappelli, P., (2000) Chambers, E., G., Foulon, M., Handfield-Jones, H., Hankin, S. M., Michaels III, E.G. (1998) Joseph, D., Ang, S., Slaughter, S. A. (2015) Ramesh, A., Gelfand, M. J. (2010)
Rewards	Employees compensation, benefits, employees' perception, and other kind of reward related variables.	Ajit, P. (2016) Allen, D. G., Bryant, P. C. and Vardaman, J. M. (2017) Cappelli, P., (2000) Chambers, E., G., Foulon, M., Handfield-Jones, H., Hankin, S. M., Michaels III, E.G. (1998) Farris, G. F. (1969) Kirschenbaum, A., Weisberg, J. (1990) Rao, T.V. (2007) Zhao, Y., Hryniewicki, M. K., Cheng, F., Fu, B., & Zhu, X. (2018)
Life Quality	Diverse variables related to life quality related to region, overtime, closeness to family and perception of stress and lifestyle respect by the company.	Ajit, P. (2016) Allen, D. G., Bryant, P. C. and Vardaman, J. M. (2017) Baek, P. (2016) Cappelli, P., (2000) Chambers, E., G., Foulon, M., Handfield-Jones, H., Hankin, S. M., Michaels III, E.G. (1998) Farris, G. F. (1969) Huffman, A. H., Adler, A. B., Dolan, C. A., Castro, C. A. (2005) Rao, T.V. (2007) Rosenbaum, E. (2019)
Leadership	Variables related to employee's leaders and the perception of the employees about their leader.	Ajit, P. (2016) Allen, D. G., Bryant, P. C. and Vardaman, J. M. (2017) Cappelli, P., (2000) Chambers, E., G., Foulon, M., Handfield-Jones, H., Hankin, S. M., Michaels III, E.G. (1998) Rao, T.V. (2007)
Job Characteristics	Characteristics of the employees' job profiles and the employee perception about it.	Ajit, P. (2016) Allen, D. G., Bryant, P. C. and Vardaman, J. M. (2017) Baek, P. (2016)

		Cappelli, P., (2000) Chambers, E., G., Foulon, M., Handfield-Jones, H., Hankin, S. M., Michaels III, E.G. (1998) Farris, G. F. (1969) Jackofsky, E. F. (1984) Kirschenbaum, A., Weisberg, J. (1990) Parasuraman and Alutto (1984) Ramesh, A., Gelfand, M. J. (2010) Rao, T.V. (2007) Spector, P. E. (1991) Zhao, Y., Hryniewicki, M. K., Cheng, F., Fu, B., & Zhu, X. (2018)
Internal Growth Opportunities	Career growth opportunities and perceptions of the employees about it.	Ajit, P. (2016) Allen, D. G., Bryant, P. C. and Vardaman, J. M. (2017) Cappelli, P., (2000) Chambers, E., G., Foulon, M., Handfield-Jones, H., Hankin, S. M., Michaels III, E.G. (1998) Farris, G. F. (1969) Kirschenbaum, A., Weisberg, J. (1990) Rosenbaum, E. (2019)
Individual Performance	Employees' performance variables.	Chambers, E., G., Foulon, M., Handfield-Jones, H., Hankin, S. M., Michaels III, E.G. (1998) Farris, G. F. (1969) Jackofsky, E. F. (1984) Zhao, Y., Hryniewicki, M. K., Cheng, F., Fu, B., & Zhu, X. (2018)
Individual Characteristics	Individual personal characteristics.	Ajit, P. (2016) Allen, D. G., Bryant, P. C. and Vardaman, J. M. (2017) Blegen, M. A., Mueller, C. W., Price, J. L. (1988) Cappelli, P., (2000) Elvira and Cohen (2001) Farris, G. F. (1969) Hom, P. W., Hulin, C. L. (1981) Kirschenbaum, A., Weisberg, J. (1990) Lee, T. W., Maurer, S. D. (1999). Rao, T.V. (2007) Zhao, Y., Hryniewicki, M. K., Cheng, F., Fu, B., & Zhu, X. (2018)
External conditions	Variables related about the labor market and the company's external image as employer.	Allen, D. G., Bryant, P. C. and Vardaman, J. M. (2017) Baek, P. (2016) Farris, G. F. (1969) Jackofsky, E. F. (1984) Kirschenbaum, A., Weisberg, J. (1990)
Education & Training	Variables related to formal education and training.	Ajit, P. (2016) Allen, D. G., Bryant, P. C. and Vardaman, J. M. (2017) Bretz, R. D., Boudreau, J. W., Judge, T.A. (1994) Farris, G. F. (1969) Rao, T.V. (2007) Zhao, Y., Hryniewicki, M. K., Cheng, F., Fu, B., & Zhu, X. (2018)
Corporate Culture, Values and Transparency	Variables related to employees' perception about the company culture, values, and expectations.	Ajit, P. (2016) Allen, D. G., Bryant, P. C. and Vardaman, J. M. (2017) Blegen, M. A., Mueller, C. W., Price, J. L. (1988) Cappelli, P., (2000) Chambers, E., G., Foulon, M., Handfield-Jones, H., Hankin, S. M., Michaels III, E.G. (1998) Erdogan, B., Bauer, T. N. (2010) Farris, G. F. (1969)

3. METHODOLOGY

The methodology used to develop the work was divided in 8 steps, based on the approach suggested for machine learning projects by Gerón (2017).

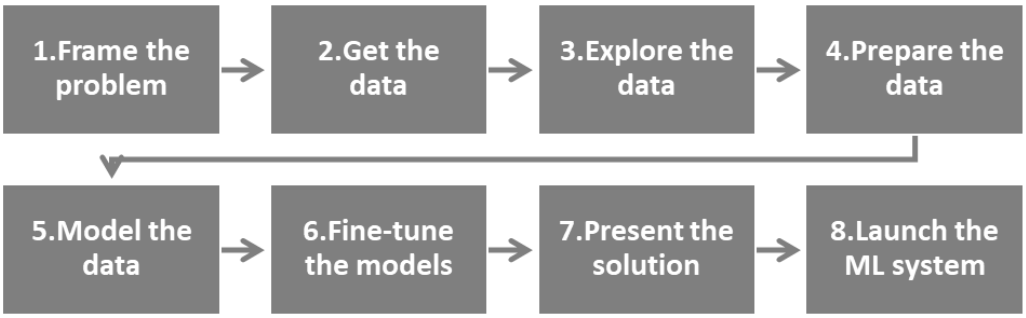

Figure 1 - Machine learning project steps

1. Frame the problem: In this first step, the problem and the study objectives are clearly defined and its broader "big picture" is understood. Important technical definitions are established such as what is the type of machine learning problem (supervised, unsupervised, etc.), if supervised what is the target variable, and what tools and techniques are going to be used.

2. Get the data: This step consists in understanding what data is needed, where it is going to be obtained, what are the data types (text, images, etc.). Also, other implications such as company approvals and legal requirements are addressed. It ends by getting the data, in the format of datasets required.

3. Explore the data: The objective of this step is to understand the data and gaining insights about it and about the problem, before the modelling. It is a moment to investigate if the assumptions made in the first step are valid. Good practices are using visualizations (e.g. box plot, histograms), which help to study features and generate insights, and document these insights to be used in the following steps.

4. Prepare the data: Preparing the data is the process of performing the transformations identified as needed to the data. It includes data cleaning (e.g. dealing with missing values, outliers, errors), creating new features, performing data normalization/standardization (which might otherwise impact models results) and feature selection to reduce the processing requirements and multicollinearity issues. Here is also the data partitioning dividing the data into training and testing datasets.

5. Model the data: Now, with the data prepared, the next step is selecting a set of models and apply them to the data. The smaller training sets can be used, also with fewer features in the beginning to evaluate error levels. Most relevant features for each algorithm should be investigated. Models should be measured and compared and, the best performing ones should be shortlisted to be better fine-tuned.

6. Fine-tune the models: At this moment, the selected models have their hyperparameters fine-tuned, and ensemble methods can be used, to look for result improvements. The best models must be fully assessed in both training and test datasets.

7. Present the solution: Here is where visualization skills are used to present the results obtained. This step should answer the problem and demonstrate the achievement or not of the objectives of the project. Even though it is mostly a communication step, proper documentation of the technical aspects must be kept.

8. Launch the ML system: This step is about getting the machine learning project ready for production within a broader production system. For example, it will be connected to a constant flow of data (e.g. from sales processes, productions processes) and generating insights into structured dashboards and being regularly retrained to improve the models. This step was not executed within this study because it was out of scope, but considerations for its implementation were given.

The steps taken during the study will be described in the following sections.

3.1. FRAME THE PROBLEM

As presented previously, the context of the turnover problem is a competitive, fast faced environment. There is a "war for talented", where skillsets are in short supply and demand is increasing, making excessive turnover a significant cost. At the same time the there is general work culture shift and new employees have different values than in previous generations, where companies need to deal with a multi-generation difficulty and many other challenges such as reducing bias for people related decisions. All of this, while attention from the leaders is required by the need of constant change to processes and systems to remain competitive, keep adding value to more demanding customers in the most efficient way they can.

Those external factors create an urging need for more professional turnover management, so that the organization influences the employee journey as a flow, providing them with the best experience and environment, enabling them to add the most value to the business. The only way of providing this experience in scale is adopting a data-driven approach, which provides the possibility of making objective decisions and create targeted and tailored strategies, which add the most value and waste less resources. As in market segmentation strategies, "mass" actions will diminish and companies will need to divide and characterize the "target markets" and "sub-markets" and take specific actions to maximize results (Kuo *et al., 2002).* This project created an initial integrated dataset and identified the most important variables that influence turnover in the studied organization. Also, it created predictive models that can be later deployed as tools for business as usual activities. Finally, with the variables selected and models trained, it identified the clusters of 'low' and, more importantly, 'high risk' individuals who are more likely for turnover, enabling potential short-term decisions and longer-term structural changes in specific processes or aspects of the employee journey.

The project utilized R (R version 3.6.2 (2019-12-12)), due to existing expertise and tool availability in the organization, allowing a smooth approval of the project in technical terms and potential easier deployment of the model into a "business as usual" tool.

3.2. GET THE DATA

The data was obtained from different sources, mainly from the companies Human Resources Management system (e.g. personal characteristics, org chart information), the performance review systems and the organizational climate survey results. The process of obtaining the data involved

obtaining the Human Resources approval and addressing the legal aspects to understand the requirements and constraints of how to deal with the data.

The dataset created includes **5722** observations, which includes active workers and workers who stopped working at the company between 01/01/2018 and 31/12/2019.

3.3. EXPLORE THE DATA

The data was explored looking directly in the database itself, using Excel, but also the *r* library *ggplot2.* Many different breakdowns and graphs were used to explore the dataset and better understand the phenomena and identify potential data issues, some of which are showed next.

The population age is mostly concentrated in the 25 to 30 years and 30 to 35 years, as seen in the histogram below (image 1).

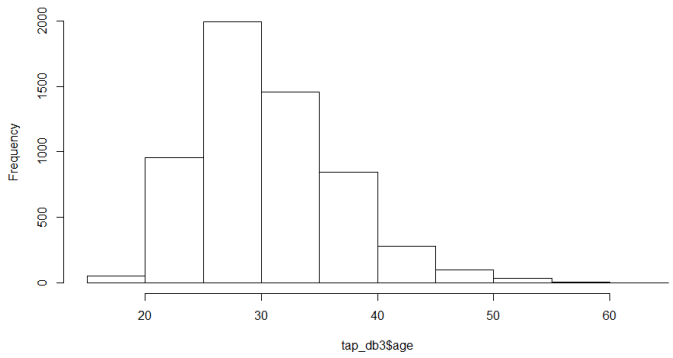

Figure 2 - Histogram of variable age

The exploration included the exploration of different variables for a basic understanding of the population characteristics, using variables such as “age” and if the individual “manages team”, as shown in the box plot next (image 2). The average of the age of individual contributors is around 30 years old, while the average age of the people who manage teams is around 35 years old.

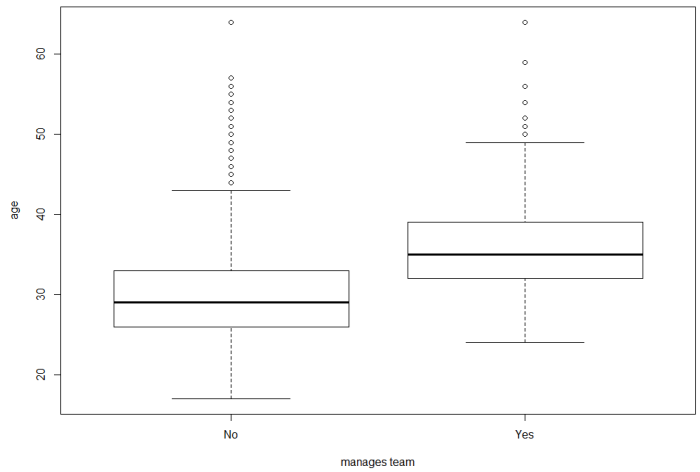

Figure 3 - Age of individual contributors and leaders

Also, the data was explored against the target variable, which is called “Out” a binary value indicating if the person’s contract with the company was terminated (considering both voluntarily and the involuntary cases). Analyzing per gender, as shown next (Image 3), no significant difference was found between males and females, even though there was a slightly higher proportion of female who left. There was a small proportion of observations without gender data, not shown in the image below.

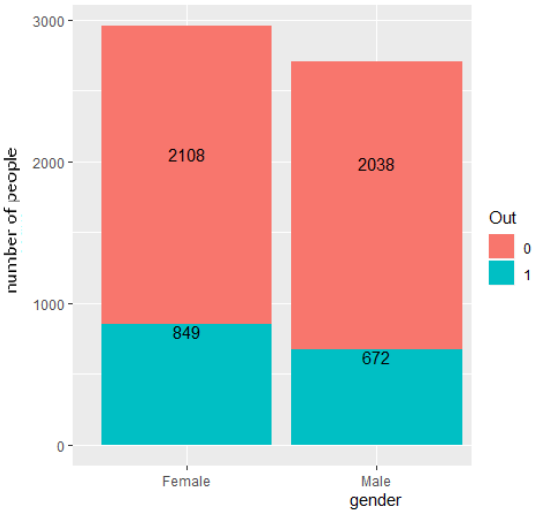

Figure 4 - Turnover by gender

The target variable “Out” was broken down by the variable “Management level”, as shown in the graph below (Image 4). Proportionally the category “supervisor” had the highest turnover, followed by the “individual contributor”. The “senior manager” category had the lowest proportion of turnover (Out = 1) in the evaluated dataset.

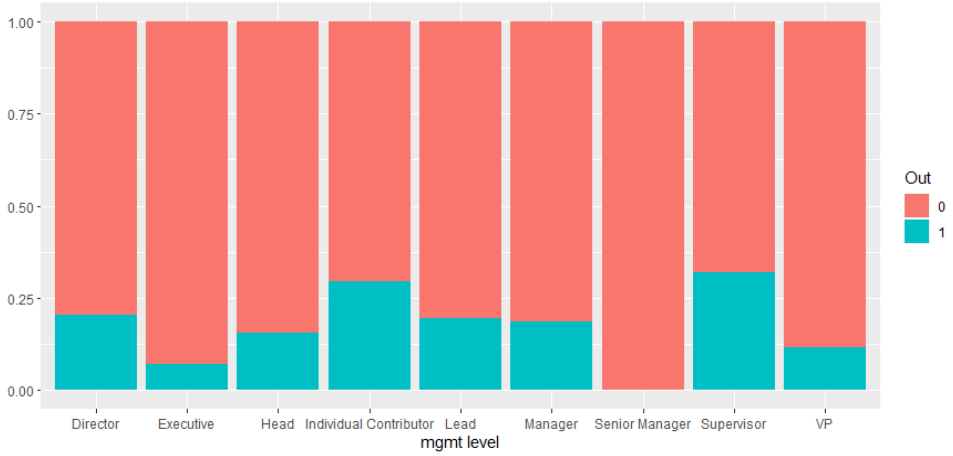

Figure 5 - Proportion of turnover by management level

When evaluating the data by “tenure” (time in the company) and “time in the job profile”, data issues were identified as there were individuals with more time in job profile then tenure, which is in practice impossible (bottom-right data points), therefore errors. The observations in the top left of the graph have longer tenure and less time in the job profile, indicating a promotion or lateral move, and the proportion of leavers (Out = 1, showed in green dots) reduces.

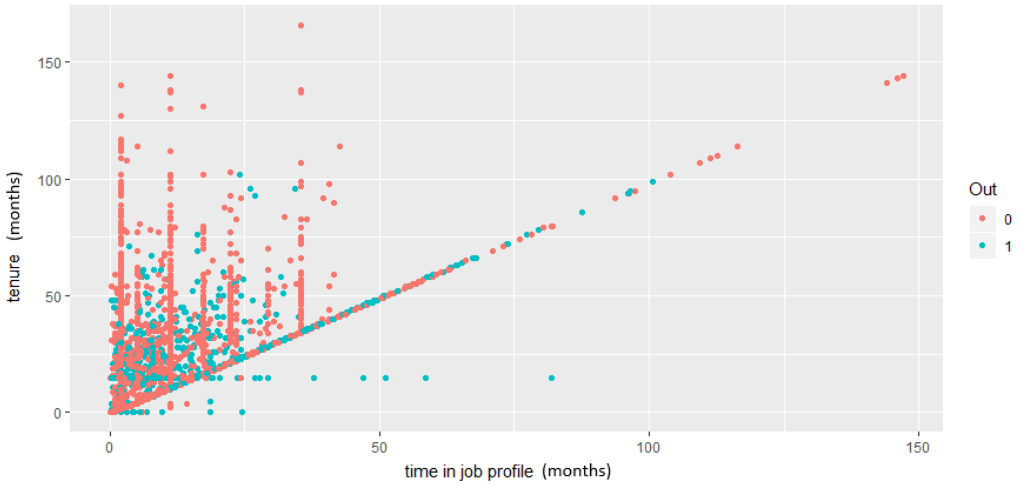

Figure 6 - Turnover by tenure (time in company) by time in the same job profile

After the exploration, all the findings were used as input for the data preparation phase.

3.4. PREPARE THE DATA

3.4.1. Feature engineering

Based on literature review, brainstorming with the human resources employees, and availability of the data within the time constraints, a dataset was created to be used as input to the model, containing a total of **46 variables** to be used as input:

Table 2 - Input variables used

Variable Group	Variable	Type	Short Description
Team Factors	wider.team.last.happiness.index	Cathegorical	Last results of a survey of the wider team (including the entire management chain) inquiring happiness.
	wider.team.last.intention.to.stay.index	Cathegorical	Last results of a survey of the wider team (including the entire management chain) inquiring intention to stay.
	wider.team.last...of.responses	Cathegorical	Last % of answers of a survey of the wider team (including the entire management chain).
	direct.peers.last.happiness.index	Cathegorical	Last results of a survey of the direct team peers inquiring happiness.
	direct.peers.last.intention.to.stay.index	Cathegorical	Last results of a survey of the direct team peers inquiring intention to stay.
	direct.peers.last...of.responses	Cathegorical	Last % of answers of a survey of the direct team peers.
	Turnover.2018	Numerical	Turnover of the area during 2018, excluding the eventual leaving of the specific employee.
	Turnover.2019	Numerical	Turnover of the area during 2019, excluding the eventual leaving of the specific employee.
	area.avg.size	Numerical	Average number of employees in the area from 01/01/2018, 01/01/2019 and 01/01/2020.
	area.avg.hires	Numerical	Average number of hires in the area from 01/01/2018, 01/01/2019 and 01/01/2020.
	area.avg.terminations	Numerical	Average number of terminations in the area from 01/01/2018, 01/01/2019 and 01/01/2020, excluding the eventual leaving of the specific employee.
Rewards	last.raise..	Numerical	The last salary raise percentage.
	long.special.leave	Cathegorical	Indicates if the worker went on a special long leave.
	short.special.leave	Cathegorical	Indicates if the worker went on a special short leave.
	time.since.pay.raise	Numerical	The time since the last salary raise happened.
Leadership	mngr.mgmt.level	Cathegorical	Identifies the management level from the worker's manager ranging from "Individual Contributor" to "Executive".
	mngr.short.special.leave	Cathegorical	Indicates if the worker's manager went on a special short leave.
	mngr.parental.leaves	Cathegorical	Indicates if the worker's manager went on parental leave.
	mngr.wfh.days	Numerical	Indicates the worker's manager's sum of "work from home" days booked in the system.
	mngr.office.location	Cathegorical	Indicates the worker's manager office Country.
	mngr.office.country	Cathegorical	Indicates the worker's manager office city.
	mngr.time.job.profile	Numerical	Indicates the worker's manager time in the current job profile.
	mngr.time.position	Numerical	Indicates the worker's manager time in the current position, which is an ID, changed only in specific contexts (transfer from countries, cost centers).
	manager.tenure	Numerical	Indicates the worker's manager time in the organization.
Job Characteristics	cost.center	Cathegorical	Cost center where the worker is allocated.
	mgmt.level	Cathegorical	Identifies the management level from the worker ranging from "Individual Contributor" to "Executive".
	manages.team	Cathegorical	Indicates if the worker has reports.
	wfh.days	Numerical	Indicates the worker's sum of "work from home" days booked in the system.
	Contract	Cathegorical	Contract type.
	same.location.mngr	Cathegorical	Indicates the worker works at the same location as the manager.
Individual Performance	X20181.performance	Cathegorical	Performance rating applied in the mid-year evaluation of 2018.
	X2018.performance	Cathegorical	Performance rating applied in the year-end evaluation of 2018.
	X20191.performance	Cathegorical	Performance rating applied in the mid-year evaluation of 2019.
Individual Characteristics	Rehire	Cathegorical	Indicates the worker had left the company and was rehired.
	Age	Numerical	Worker's age.
	Generation	Cathegorical	Worker generation category (e.g. Millennial, Baby Boomer).
	Gender	Cathegorical	Worker's gender.
	Dependents	Numerical	Number of dependents.
	parental.leave	Cathegorical	Indicates if the worker went on parental leave.
	time.in.job.profile	Numerical	Indicates the worker's time in the current job profile.
	time.in.position	Numerical	Indicates the worker's time in the current position, which is an ID, changed only in specific contexts (transfer from countries, cost centers).
	Nationality	Cathegorical	Nationality of the worker.
	2nd.nationality	Cathegorical	Indicates if the worker has or not a second nationality.
	Tenure	Numerical	Indicates the worker's time in the organization.
External conditions	office.city	Cathegorical	Worker's office city.
	office.country	Cathegorical	Worker's office country.

A lot of the variables were created during the project and, even the ones that already existed, were mostly never analyzed together.

Also, other data adjustments were made when testing algorithms, for example the "nationality" variable was adapted to reduce the number of classes, since the RandomForest algorithm did not support a variable with more than 52 options of classes. The Neural Network algorithm also required feature engineering to turn categorical variables into dummy variables, since all the input to this algorithm needs to be quantitative in the respective *r* package.

3.4.2. Data cleaning

Data cleaning is a process of applying tools and performing tasks in order to handle situations that affect the accuracy of the data, such as missing values, noise and inconsistencies, it is a complex effort (Han *et al*., 2011).

For missing values in categorical variables, a 'Not Available' class was created, for the cases in numerical variables the criteria chosen to replace them was the median. Values that were manually identified as outliers were also replaced by the median.

A total of **1491** data features were included or corrected, **0,57%** out of a total of **263212**, affecting **728** observations, **13%** of all observations.

Normalization was the following talk. It is important because many machine learning algorithms are sensitive to different scales of evaluation. The min/max approach was used in the **17 numerical variables**. This method was chosen because the dataset distribution was not linked to a normal distribution. The min/max method is sensitive to outliers making it crucial the previous steps or understanding and correcting potential mistakes in outliers. The dataset was called tap_db3 and the example below shows the code used in R for the variable "age".

```
tap_db3$age <- (tap_db3$age - min(tap_db3$age))/(max(tap_db3$age)-min(tap_db3$age))
```

3.4.3. Feature selection

For feature selection two techniques were used. The first was the stepwise, aiming to maximize entropy with the least possible variables. Then a correlation matrix was used to evaluate further opportunities of reduction.

The stepwise approach is a method that performs incremental changes in the variables and tests, adding or removing features, and evaluating the impact in the data using the Akaike Information Criterion (AIC). This metric evaluates the information loss of the model when removing features and gives penalties if there are more features in the model.

There are three methods to apply the stepwise: 'backward', that starts with all the features and removes one by one; 'forward' that starts with one feature and adds one by one; and 'both' that starts with all features which removes one by one but also tests adding back variables.

The three approaches were used and had slightly different outputs. The variables that were selected in the three stepwise models were maintained. Afterwards a correlation matrix was created, to evaluate further potential for reducing the variables.

The libraries used were *glmnet* for the stepwise process and *corrplot* to plot the correlation matrix.

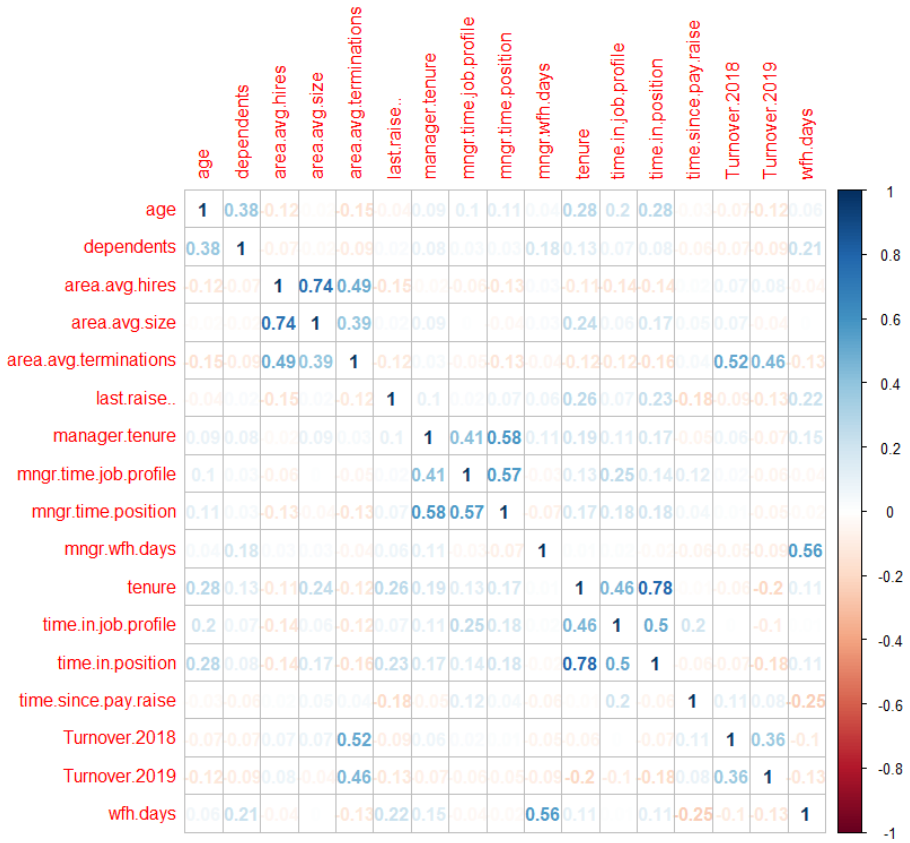

Figure 7 - Correlation matrix

Base on the Pearson correlation coefficients, it was not identified any relation usually classified as "very strong", with values above 0,8 or below -0,8. However, two set of variables had "strong" correlations (above 0,6): the highest relation identified was between 'tenure' and 'time.in.job.position' (0,78), but the 'tenure' variable was already removed during the previous stepwise techniques. The second strongest relation was between, 'area.avg.size' and 'area.avg.hires' (0,74), out of which the variable 'area.avg.hires' was removed after an evaluation of the variable impact using the random forest algorithm.

In the end of the process, 26 variables were prioritized to train the models :

- area.avg.size
- area.avg.terminations
- contract
- cost.center
- direct.peers.last...of.responses
- direct.peers.last.happiness.index
- last.raise..

- manager.tenure
- manages.team
- mgmt.level
- mngr.time.job.profile
- mngr.wfh.days
- nationality
- office.city
- parental.leave
- rehire
- same.location.mngr
- time.in.job.profile
- time.in.position
- time.since.pay.raise
- Turnover.2019
- wfh.days
- wider.team.last...of.responses
- X2018.performance
- X20181.performance
- X20191.performance

The target variable was called “Out” and it identifies if the person stopped working at the company in the studied period 01/10/2018 and 31/12/2019, returning a 1 if positive, 0 if negative.

3.4.4. Data partitioning

The dataset was divided into two subsets: 70% as training set, 30% as testing/validation set, as showed below in table 3:

Table 3 - Data partitioning

Dataset	Total
Train	4030
Test	1692
Total	5722

Burkov (2019) explains that data partitioning aims to avoid overfitting. If the model was trained with 100% of the data it would reflect perfectly its behavior losing its prediction capacity and, therefore, probably performing poorly with new data. That is why the data is divided in the training set, usually with 70% of the data to train the model, and 30% to validate the results. While the training set is used to adjust the models’ parameters during the model training, the testing set is not, and is only used to evaluate the outcomes of the model, for example in terms of accuracy. The models that best predict the test set results were considered the best models.

3.5. MODEL THE DATA AND FINE-TUNE THE MODELS

The following steps included modeling the data and fine tuning the models, steps five and six in the methodology utilized. Gerón (2017) describes the "model the data" step as the process of selecting models and applying them to the data available. All the different models applied should be measured and compared, against the train and test datasets. Afterwards, the following step of "fine tuning the model", includes adjusting hyperparameters, identifying the best performing algorithm and performing the complete assessment in train and validation sets. The models selected were random forest, logistic regression and neural networks.

3.5.1. Random Forest

Random Forest is an ensemble method of classification that produces a group of decision trees and, based on their output, delivers a classification or also regression (e.g. mean of the predictions) of the individual trees.

Each decision tree is made of nodes and leaves, where a node is a specific 'decision' of the tree and the "leaves" are the specific classifications. The objective of a decision tree is to discriminate between classes, obtaining leaves that are as pure as possible where, ideally, each leaf will only contain individuals one specific class. In the illustrative example below (Figure 8) the target variable is 'ideal exercise', and the input variables are 'Prefers indoors activities', 'is in shape' and 'prefers team activities'.

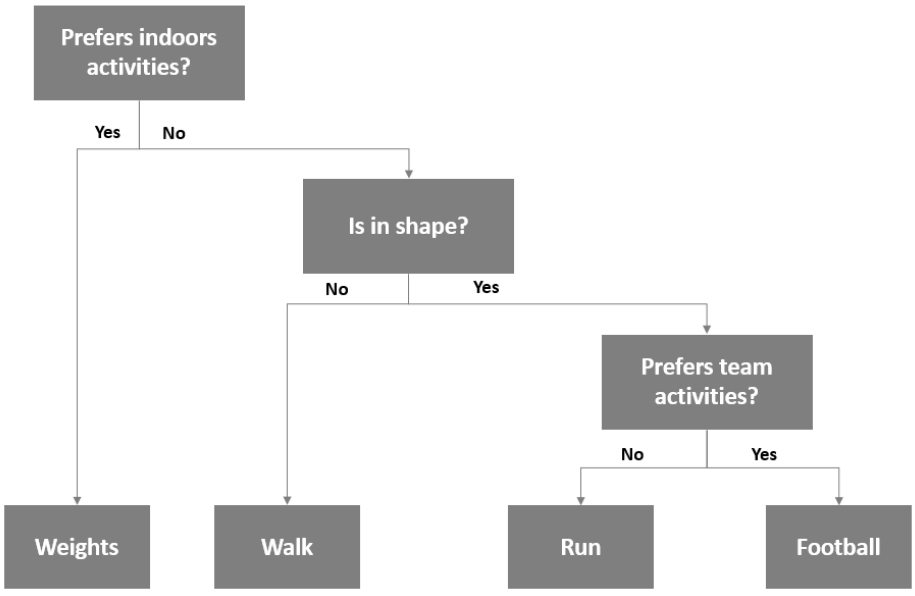

Figure 8 - "Ideal exercise" decision tree example - created by the author

Decision trees have the advantage of having an easier interpretation, they also have no problem dealing with varied data types and are not sensible to scale factors. In the Random Forest approach, the advantage of easier interpretation is lost since the outcome is a result of a vote (or regression) of the individual trees.

In the individual decision tree approach, there is an objective of choosing the variables which reduce the node impurity the most, measured by the Gini Importance, therefore testing all variables in each node, being the measure how well the data is divided after that node or decision is made.

In the Random Forest the randomness when building the trees and training the model is desirable, since variance between the trees is more likely to produce better outputs of the global "forest".

The randomness is introduced by choosing a random training set for each base learner and by the choosing the nodes of each the decision tree from a random subset of attributes of the training set.

A measure that evaluates the importance of each of the input variables related to the target variable is the Mean Decrease in Gini. It measures the average (mean) of a variable's total decrease in node impurity, weighted by the proportion of samples reaching that node in each individual decision tree in the random forest. That way, it measures how useful a variable is for estimating the target variable over the group of the trees that are part of the forest.

The library *randomForest* was used in R. The hyperparameters configured were the following:

- Number of trees 500
- Number of variables tried at each split: 5

The mean decrease Gini graph below (Figure 9) shows the variable importance calculated from the model trained according to the Random Forest algorithm.

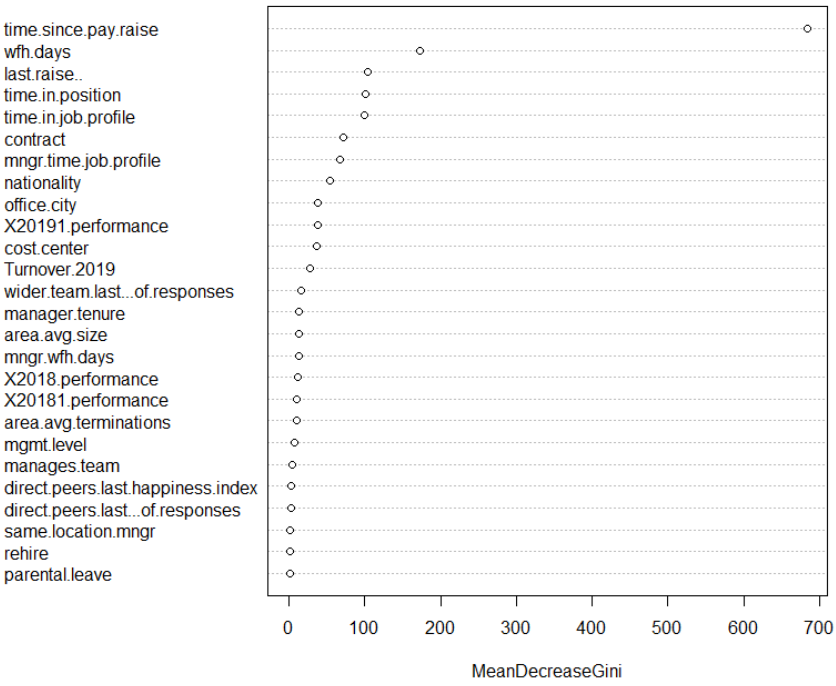

Figure 9 - Variables importance according to random forest algorithm

3.5.2. Logistic Regression

Logistic regression is an algorithm that performs a classification, not a regression. It is a generalized linear model (GLM) with a binomial random component. Unlike the linear regression which returns as output a continuous numerical value, the output of the logistic regression is a binary output of either 0 or 1, or as a probability of being 1 or 0.

In order to obtain the value of the parameters, the maximum likelihood estimation is used. It needs to establish the better predictors that maximize the probability (likelihood) of observing the output. These are iterative algorithms, starting with arbitrary values and repeating until the log-likelihoods are not changed significantly.

The input dimensions are organizer in a linear manner, but unlike the linear regression, the output is fed into a logistic function which transforms it into a nonlinear output, which is interpreted as the probability of the input belonging to class 1 (in the interval [0,1]). After that, a transformation is usually performed estimating that if probability *p* is higher than 0,5 than the output should be 1, otherwise it should be 0.

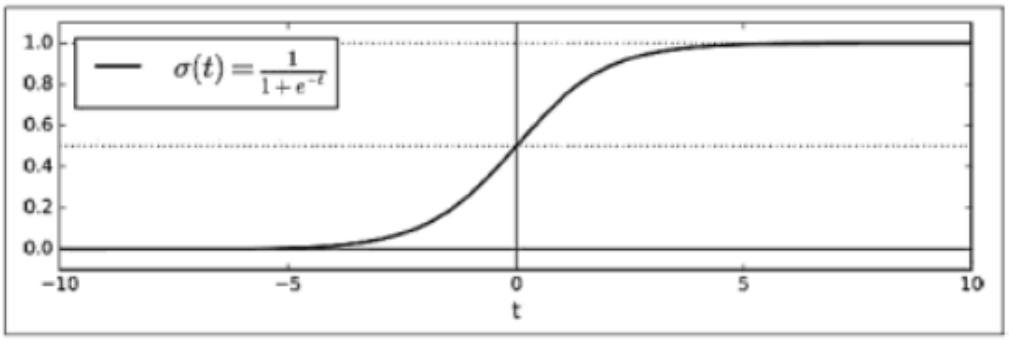

Figure 10 - Logistic function (Gerón, 2017)

Its main advantage in comparison with other classification methods is that it allows a clearer interpretation of the results, delivering a “formula”, with clear positive and negative impacts from the input variables. In linear regression the coefficients can be directly be understood in terms of impact in the target, but in the logistic regression, on the other hand, that interpretation cannot be made directly, because the output is the probability. Therefore, the one interpretation that can be made is of a positive impact (sign that the curve ascends), and vice-versa.

In R, the library *glmnet* was used. The model translates all the categorical variables into multiple binary variables to use them as input, so from the 26 variables the input ended being turned into 106 variables. Out of those, the model identified the following 11 variables (including binary variable categories turned into separate variables) with highest impact on the target variable “Out”, listed in table 4 below:

Table 4 - Most impactful variables according to Logistic Regression

Variable/Variable Category	Direction of Impact in Target Variable	Absolute Impact Relative Position
time.since.pay.raise	+	1
time.in.job.profile	-	2
mngr.time.job.profile	-	3
area.avg.size	+	4
nationalityNot Available	-	5
contractNot Available	+	6
wider.team.last...of.responsesNot Available	+	7
Turnover.2019	+	8
last.raise..	+	9
time.in.position	-	10
wfh.days	+	11

An important highlight to be given is the potential limitation of the logistic regression. Since it is a linear model, it will not capture non-linear behavior of input variables.

3.5.3. Neural Network

A neural network is an algorithm vaguely inspired in the functioning of the human brain. (Gerón, 2017). It receives input from the environment or previous 'neuron cells', which is fed through connections with 'weights' (synapses), producing a weighted sum from the sources. This value, as in the logistic regression, is fed into an activation function (non-linear such as the sigmoid), that converts the input into output, which goes to the next neuron or the environment.

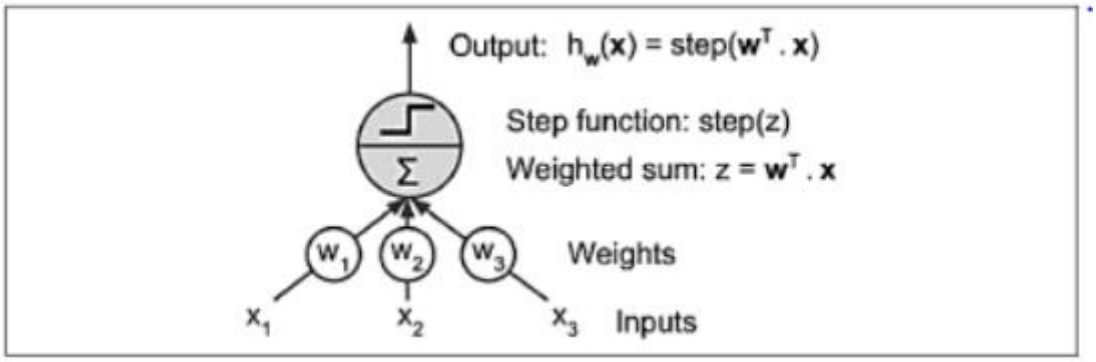

Figure 11 - Perceptron, the simplest artificial neural network structure (Gerón, 2017)

During the modelling the training algorithm will adjust the weights in the 'synapses' so that the error of the output is reduced. It focuses on many small adjustments instead of big changes. This iterative process of evaluating error and adjusting the weights is what is called backpropagation.

The advantages of the algorithm are that it can be applicable to many kinds of data and complex problems and is especially useful when little is known about the phenomena. However, its main

disadvantages are that the outputs may be of difficult interpretation and the processing performance can be worse that other algorithms.

The r package *neuralnet* was used to train the model studied and as hyperparameters the number of neurons was evaluated from 4 to 9.

3.5.4. Models results

The algorithms that were modeled, using the training set and their prediction capacities were evaluated against the test dataset. The confusion matrix, as shown in Figure 12 below, is a simple and useful way to represent it (Burkov, 2019). Each of the four quadrants groups the observations positioned according to the evaluation of the value of the target variable, classified in the intersection of the classes predicted (true x false) and actual values (true x false).

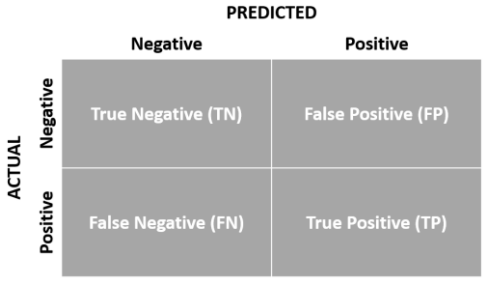

Figure 12 - Confusion matrix quadrants, created by the author

The main metrics available based on the matrix are:

- the true negative (TN), where the predicted and actual values are negative;
- the true positive (TP), where the predicted and actual values are positive;
- the false positive (FP), where the prediction was positive, but the actual value was negative;
- the false negative (FN), where the prediction was negative-, but the actual was positive;
- Precision, TP / (TP+FP). Out of all positive predictions, how often is it correct;
- Recall, TP/ (TP+FN). Out of all actual positives, how much was correctly predicted;
- Misclassification Rate (FP+FN)/total;
- Accuracy (TP+TN)/total.

The metric chosen for the model evaluation was the misclassification rate, calculated for both the training and test datasets, for each of the algorithms, as seen in Table 5.

Table 5 - Model evaluation

Model	Misclassification Rate Training Dataset	Misclassification Rate Test Dataset
Random Forest	1,04%	0,65%
Logistic Regression	3,47%	4,31%
Neural Network (4 neurons)	0,20%	6,74%
Neural Network (5 neurons)	0,00%	6,97%
Neural Network (6 neurons)	0,00%	7,21%
Neural Network (7 neurons)	0,00%	6,62%
Neural Network (8 neurons)	0,00%	5,44%
Neural Network (9 neurons)	0,00%	6,80%

This was an iterative process, and data had to be adjusted (e.g. reduction of possible of classes of variables) as errors and issues were perceived when running the code and packages in *r*.

3.6. PRESENT THE SOLUTION

After identifying the most relevant variables and creating the classification predictive model, the following step adopted was to divide the dataset in clusters for a qualitative evaluation to generate the final insights. This is a clustering machine learning problem, which consists in assigning a label to observations by leveraging an unlabeled dataset (Burkov, 2019). The K-means method was used in this project. Kuo *et al.* mentions the non-hierarchical methods, applied in similar classification problems (client segmentation, for example), using methods such as K-means, can provide higher accuracy, if the initial number of clusters is properly selected.

The initial activity in the K-means method was choosing the number of clusters, then randomly putting k feature vectors called centroids (1 per cluster). After that, the algorithm computes the distance from each observation to each centroid and, finally, labels the observations to the cluster which minimizes the sum of square errors within the clusters, in other words, minimizing the distances from the observations from the centroids. The variables used for clustering were the same 26 variables selected during the stepwise process and the correlation matrix.

For determining the number of clusters, a technique called “elbow graph” was used. Figure 13 shows the decrease of the total error as we increase the number of clusters, showing a significant reduction from three to four clusters, and a smaller decrease after that. Therefore, the number of clusters chosen was four. The *r* library *cluster* was used to apply these techniques.

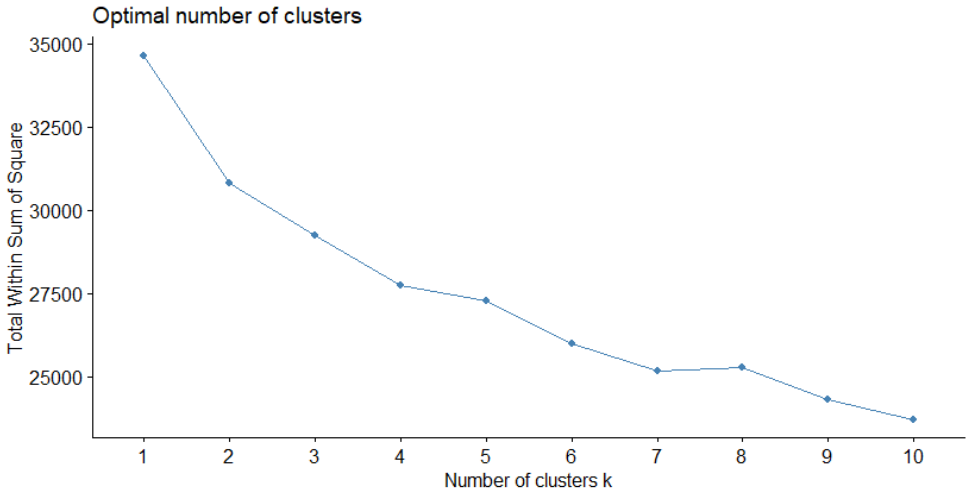

Figure 13 - Optimal number of clusters

The distribution of observations by cluster is shown below in Table 6.

Table 6 - Cluster sizes

Cluster	# of observations	% of observations
1	1620	28%
2	905	16%
3	1781	31%
4	1416	25%
TOTAL	**5722**	**100%**

The next step was to evaluate each cluster in terms of employee turnover, using the "Out" variable and, next, to describe the characteristics of the clusters that showed more extreme behaviors. Table 7 below shows the clusters and the respective turnover rate.

Table 7 - Turnover by cluster

Cluster	a.Active employees in 01.2020	b.Employees "Out" from 01.18 to 01.20	c.Turnover *(b/(a+b))*
1	1445	175	10,8%
2	724	181	20,0%
3	942	839	47,1%
4	1040	376	26,6%
Total	**4151**	**1571**	**27,5%**

Clusters 1 and 3 presented the more extreme behaviors, with turnovers more divergent from the global average (27,5%). In order to perform a more detailed analysis, the variables used in the clustering process were analyzed, making comparisons between the four clusters. Table 8, below, shows a selection of the variables that presented extreme behavior in those two clusters (1, 3 or both) in relation to the global average of the population, being either the highest or the lowest value. A color scale, used from the second until the fifth column of the table from the left to the right,

indicates the relative distance of the observation within a cluster compared with the population average, being greener the values lower than the average and redder the values higher than the average. The two columns to the right indicate the relative position of the first and third clusters, where those clusters presented the more extreme behaviors.

Table 8 - Cluster comparison

	Cluster results compared to the global average				Cluster behaviour class	
VARIABLE	1	2	3	4	1	3
Average of time since pay raise	-28%	-6%	37%	-11%	Lowest	Highest
Average of last raise percentual increase	50%	26%	-40%	-24%	Highest	Lowest
Average of time in position (code)	38%	38%	-47%	-9%	Highest	Lowest
Average of 'work from home' days	62%	2%	-31%	-33%	Highest	
Average of manager 'work from home' days	42%	-21%	-3%	-31%	Highest	
Average of team turnover in 2019	-33%	-36%	47%	3%		Highest
Average of manager time in current job profile	-7%	45%	-10%	-7%		Lowest
Contract - Category 'Permanent'	39%	30%	-98%	60%		Lowest
Performance 2018.1 - Category 'Not Available'	-48%	-7%	29%	23%	Lowest	Highest
Performance 2018.1 - Categories of 'High Performers'	119%	49%	-89%	-57%	Highest	Lowest
Performance 2018 - Category 'Not Available'	-78%	-17%	52%	35%	Lowest	Highest
Performance 2018 - Categories of 'High Performers'	130%	51%	-93%	-65%	Highest	Lowest
Performance 2019.1 - Category 'Not Available'	-85%	-13%	51%	43%	Lowest	Highest
Performance 2019.1 - Categories of 'High Performers'	152%	18%	-84%	-80%	Highest	Lowest
Wider team climate survey last % of responses - Category 'Very High'	62%	2%	-49%	-11%	Highest	Lowest
Managers Team - Category 'Yes'	-25%	200%	-81%	2%		Lowest
Parental Leave - Category 'Yes'	54%	42%	-78%	10%	Highest	Lowest
Same location as manager - Category 'Yes'	2%	-6%	3%	-1%		Highest

Cluster 1 has 10,8% turnover, so was named the *Loyal* cluster. Based on the analysis, it can be observed that the employees in this cluster, in comparison with the others, have: received a pay raise more recently and of higher percentage, have been in the company for a longer period, have used more ‘work from home’ days as have their managers, have much lower ‘not available’ data in performance, have higher concentration of the top performers, have the highest participation on climate surveys and have taken parental leave more frequently.

Cluster 3 has 47,1% turnover, so was named *Flight Risk* cluster. The employees classified in this cluster were people who: have not had a raise for the longest period and have received the lowest raises, have been in the company for shorter period, are from teams with more team turnover recently, have managers with lowest time as managers in the company, have fewer ‘permanent’ work contracts, have not formally received their performance ratings, having highest ‘not available’ data and having the lowest incidence of ‘high performers’, have lowest incidence of “very high” percentual of responses in the climate survey, mostly do not have teams managed by them, have not taken parental leave, they also have their managers mostly working in the same location as them.

3.7. LAUNCH OF THE ML SYSTEM

With the model created, the next step would be to launch it as part of a broader machine learning system, connected with actual business decisions. It was not the scope of this work to complete this step, however some implementation considerations should be considered.

The creation of the model is a process that involves iterations, to constantly improve data quality, correct issues, create new variables, but it is not useful to the organization if it is not generating action. Allen *et al* (2017) divides retention actions in systemic, which result in general principles and rules for the broad organization, and targeted, which are based on specific drivers and often influence turnover among a specific group of employees. The work developed can provide input to both, but specially the last results presented, with the clusters descriptions can provide very clear insights for the target actions.

The next step is turning this model, with its structured datasets and *r* code into a business as usual tool, that generates insights regularly. Its outcomes can either motivate the systemic actions, such as the change of a global policy of compensation, if it is identified as a key variable, for example, or targeted actions for specific groups of individuals (such as the high risk ones identified).

For this to happen, it is necessary to discuss the initial insights obtained during the work and create the conditions for the data to be generated in an automatic and reliable manner. That way it would be possible for the insights to be discussed regularly by the relevant stakeholders, so the data and the model would be constantly improved, and the decisions made would become more data-driven and possible constantly re-evaluated, in recurring improvement cycles. This improvement cycles would allow the organization to focus on concrete actions, tackling either on specific features of the organization (e.g. specific policies) of specific sub-set of individuals, generating reduction in unwanted turnover with the lowest waste of resources possible.

4. Conclusions

In a context of fast-paced changes, where talents and specific skills become scarcer, the competition becomes constantly fiercer. With the increasing difficulties and costs of hiring and onboarding new talent and replacing existing talents, managing the turnover "flow" is key to ensure business continuity and the achievement of its objectives. Actions to understand and maintain employee turnover in desired levels are a huge challenge and, in order to be effective, the plans should involve identifying what features predict the employee turnover, using quantitative techniques, for data-driven targeted strategies to be designed.

This project was developed in a company deeply inserted in this scenario and aimed to apply the tools of advanced predictive analytics in order to support the organization to solve the problem of identifying who are the employees that have the highest and lowest turnover probability.

Following the method suggested by Gerón (2017), the machine learning project was developed to solve the initial problem stated. Before 'deep diving' into the data, the problem was better framed, and the literature review of the turnover phenomena gave broader understanding of the context and several potentially relevant variables were identified.

After that, the employee data was gathered from the studied company, an integrated dataset was created and the data was explored. The preparation of the data included dealing with missing values, outliers, normalization of the values and ended with the feature selection. With the stepwise method and correlation matrix, the 26 most relevant variables impacting turnover were prioritized.

The dataset, with the prioritized variables, was partitioned and the data was modeled using three different supervised machine learning algorithms to predict turnover: Random Forest, Logistic Regression, Neural Networks. The best prediction model resulted from the Random Forest algorithm which had a misclassification rate of only 0,65% in the test dataset.

In order to present a solution to the initial problem, the K-means unsupervised machine learning clustering algorithm was applied on the employee dataset resulting in four clusters, two of which had more extreme turnover behaviors. Cluster 1, the 'Loyal', had a low turnover rate (10,8% against a population average of 27,5%) and Cluster 3, the 'Flight risk', with a turnover rate (47,1%), much higher than the average. Both clusters were characterized according to the variables in which they showed more extreme behavior, providing further insights that give a 'clearer picture' of the people within those groups.

After the creation of the turnover predictive model and the cluster analysis, the challenges can be divided in two: launching these as machine learning 'live' models in the company, which are constantly updated and improved as the business reality changes; and turning insights into concrete actions that improve the human resources policies, processes, procedures and, in summary, generate data-driven, more targeted actions that improve the employee experience in the company.

This study delivers analytics insights and tools that can be the basis for the forementioned targeted strategies and, if implemented, could allow the company to better influence the turnover to behave as the 'river flow' desired. Hopefully it will, consequently, help leader to make decisions that provide a better professional journey for the employees and deliver more value for the business.

5. Limitations and recommendations for future works

Future works can explore more specific kinds of turnover, setting as target variable only voluntary or only involuntary turnovers. Another possibility is exploring as target variable a specific time frame such as a binary variable if left the company before 1 year or 6 months, to give insight more clearly situated in time. These possibilities would allow to create a narrower focus that might help to create more actionable insights.

Another limitation was regarding the data availability. Since it was the first effort, data was scattered and there was little understanding of the use of this kind of analysis. Future works could aggregate more data, such as from withdrawal processes, such as *LinkedIn* data or other metrics not explored in this work, such as the compensation ones.

Other potential improvement is to include variables from work outcomes metrics, aggregating other dimensions different from the typical HR data, providing a more complete dimension, especially for those areas and roles that have more direct metrics (e.g. commercial, operations, support).

A general limitation was regarding the history of the data, as processes and systems utilized were relatively new, and even the ones in placed changed (e.g. change in categories), this decreased the data quality. An important recommendation is to avoid unnecessary changes or, at least, keep a clear track of those, so data does not lose its integrity or it can be reconciled and can be used to achieve better results.

Another important issue to be considered are the ethical concerns and challenges emerging. Chamorro-Premuzic and Bailie (2020), raise this exemplifying the latest technological advances that are happening such as wearable gadgets, which might create opportunities, but also major privacy concerns. The exploration of people analytics, as a whole, needs to be clearly set within an ethical framework not to break employee trust and result in a poorer experience for employees and worse results for the organization.

6. Bibliography

Agovino, T. (2019, February 23rd). To have and to Hold. Retrieved in December 22, 2020 from: https://www.shrm.org/hr-today/news/all-things-work/pages/to-have-and-to-hold.aspx

Ajit, P. (2016). Prediction of employee turnover in organizations using machine learning algorithms. algorithms. (IJARAI) International Journal of Advanced Research in Artificial Intelligence, Vol. 5, No. 9, 2016.

Allen, D. G., Bryant, P. C. and Vardaman, J. M. (2017, November 30). Retaining Talent: Replacing Misconceptions with Evidence-Based Strategies. Academy of Management Perspectives Vol. 24, No. 2

Baek, P. (2016). New trends in people analytics. Retrieved in November 15, 2019 from: http://digitalcommons.ilr.cornell.edu/student/141

Bassi, L., McMurrer, D. (2007). Maximizing Your Return on People. Harvard Business Review, March 2007

Becker, S. (November 5th, 2019). Showcase for a data-driven HR: Understand the business impact of employee's attrition. Retrieved in December 15, 2020 from: https://www.linkedin.com/pulse/showcase-data-driven-hr-understand-business-impact-employees-becker/

Blegen, M. A., Mueller, C. W., Price, J. L. (1988). Measurement of kinship responsibility for organizational research. Journal of Applied Psycology, 73, 402-409.

Bretz, R. D., Boudreau, J. W., Judge, T.A. (1994). Job search behaviour of employed managers. Personnel Psychology, 47, 275-301.

Burkov, A. (2019): The Hundred-page Machine Learning Book. 1st edition. Andriy Burkov.

Cascio, W. F. (2006). Managing human resources: Productivity, quality of work life, profits (7th ed.). Burr Ridge, IL: Irwin/McGraw-Hill

Cappelli, P., (2000). A Market-Driven Approach to Retaining Talent. HBR January-February 2000 Issue.

Chambers, E., G., Foulon, M., Handfield-Jones, H., Hankin, S. M., Michaels III, E.G. (1998). The war for talent. The Mckinsey Quarterly, volume number 3.

Chamorro-Premuzic, T., Bailie, I. (2020, October 21st). Tech is transforming people analytics. Is that a good thing? Retrieved in January 10, 2021 from: https://hbr.org/2020/10/tech-is-transforming-people-analytics-is-that-a-good-thing

Chen, Y., Lin, Y., Kung, C., Chung, M., Yen, I. (2019). Design and Implementation of Cloud Analytics-Assisted Smart Power Meters Considering Advanced Artificial Intelligence as Edge Analytics in Demand-Side Management for Smart Homes. *Sensors. Volume 19 (9): pp.2047.*

Fan, C. Y., Fan, P. S., Chan, T. Y., & Chang, S. H. (2012). Using hybrid data mining and machine learning clustering analysis to predict the turnover rate for technology professionals. Expert Systems with Applications, 39(10), 8844-8851.

Erdogan, B., Bauer, T. N. (2010). Differentiated leader-member exchanges: The buffering role of justice climate. Journal of Applied Psychology, 95, 1104-1120.

Elvira, M. M., Cohen, L. E. (2001). Location matters: A cross-level analysis of the effects of organizational sex composition on turnover. Academy of Management Journal, 44, 591-605.

Farris, G. F. (1969, October). A predictive study for turnover. Massachusetts Institute of Technology.

Gerón, A. (2017): Hands-on Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, 1st edition. O'Reilly Media.

Guenole, N., Ferrar, J., Feinzig, S. (2017)- The Power of People: Learn How Successful Organizations Use Workforce Analytics To Improve Business Performance. Pearson FT Press, 1st edition.

Han, J., Pei, J., Kamber, M. (2011). Data Mining: Concepts and Techniques. The Morgan Kaufmann Series in Data Management Systems. 3rd Edition.

Hausknecht, J. P., Hom, P. W., Lee, T. W., Shaw, J. D. (2017). One Hundred Years of Employee Turnover Theory and Research. Journal of Applied Psychology. Vol. 102, No.3, pp. 530-545.

Hom, P. W., Hulin, C. L. (1981). A competitive test of the prediction of reenlistment by several models. Journal of Applied Psychology, 66, 23-39.

Huffman, A. H., Adler, A. B., Dolan, C. A., Castro, C. A. (2005). The impact of operations tempo on turnover intentions of Army personnel. Military Psychology, 17, 175-202.

Jain, P. (2015, December 12th). Top 5 reasons why analytics projects fail. Retrieved in November 30, 2020 from: https://www.forbes.com/sites/piyankajain/2015/12/12/5-reasons-why-analytics-projects-fail/?sh=3443b55b6507

Jackofsky, E. F. (1984). Turnover and Job Performance: An Integrated Process Model. Academy of Management Review, Vol. 9, No 1, pp. 74-83.

Joseph, D., Ang, S., Slaughter, S. A. (2015). Turnover of turnaway? Competing risk analysis of male and female IT professionals' job mobility and relative pay gap. Information Systems Research, 26, 145-164.

Kirschenbaum, A., Weisberg, J. (1990). Predicting Worker Turnover: An Assessment of Intent on Actual Separations. Human Relations, Volume 43, Number 9, pp. 829-847.

Kuo, R. J., Ho, L. M., & Hu, C. M. (2002). Integration of self-organizing feature map and K-means algorithm for market segmentation. Computers & Operations Research, 29(11), 1475–1493.

Ledet, E., McNulty, K., Morales, D, Shandell, M. (2020, October 2nd). How to be great at people analytics. Retrieved in November 19, 2020 from: https://www.mckinsey.com/business-functions/organization/our-insights/how-to-be-great-at-people-analytics

Lee, T. W., Maurer, S. D. (1999). The effects of family structure on organizational commitment, intention to leave and voluntary turnover. Journal of Managerial Issues, 11, 493-513.

Morrison, R. (2018, December 18th). Rethinking workforce planning for a disruptive age: OP&A leads the way. Retrieved in December 20, 2020 from: https://www.orgvue.com/resources/articles/rethinking-workforce-planning-for-a-disruptive-age-opa-leads-the-way/

Parasuraman, S., Alutto, J. A. (1984). Sources and outcomes of stress in organizational settings: Toward the development of a structural model. Academy of Management Journal, 27, 330-350.

Ramesh, A., Gelfand, M. J. (2010). Will they stay or will they go? The role of job embeddedness in predicting turnover in individualistic and collectivistic cultures. Journal of Applied Psycology, 95, 807-823.

Rao, T.V. (2007, May 1). Factors affecting attrition and strategies of retention. NHRD Journal.

Rosenbaum, E. (2019, April 3rd). IBM artificial intelligence can predict with 95% accuracy which workers are about to quit their jobs. Retrieved in January 19, 2020 from: https://www.cnbc.com/2019/04/03/ibm-ai-can-predict-with-95-percent-accuracy-which-employees-will-quit.html

Rubenstein, A.A., Eberly, M.B., Lee, T.W., Mitchell, T.R. (2018). Surveying the forest: A meta-analysis, moderator investigation, and future-oriented discussion of the antecedents of voluntary employee turnover. Personnel Psycology, 2018; Volume 71, pp 23-65.

Schweyer, A. (2018). Predictive Analytics and Artificial Intelligence in People Management. Retrieved in November 2019 from Incentive Research Foundation from IRF site: irf.org

Spector, P. E. (1991). Confirmatory test of a turnover model utilizing multiple data sources. Human Performance, 4, 221-230.

Zhao, Y., Hryniewicki, M. K., Cheng, F., Fu, B., & Zhu, X. (2018, September). Employee turnover prediction with machine learning: A reliable approach. In Proceedings of SAI intelligent systems conference (pp. 737-758). Springer, Cham.

7. Appendix

Table 9 - Literature review of variables impacting turnover

Variable Group	Variable Standardized	Authors
Work Environment	Lunch table design	Baek, P. (2016)
	Natural light (number of windows)	Baek, P. (2016)
	Physically working next to others	Baek, P. (2016)
	Places occupied during the day (heatmaps)	Baek, P. (2016)
	Perception of working conditions	Ajit, P. (2016)
	Perception of working environment	Cappelli, P., (2000)
Withdrawal Process	Job Search Behaviours	Bretz, R. D., Boudreau, J. W., Judge, T.A. (1994)
		Allen, D. G., Bryant, P. C. and Vardaman, J. M. (2017)
		Schweyer, A. (2018)
	Job Search Intentions	Allen, D. G., Bryant, P. C. and Vardaman, J. M. (2017)
	Intention to leave	Allen, D. G., Bryant, P. C. and Vardaman, J. M. (2017)
		Kirschenbaum, A., Weisberg, J. (1990)
Team Factors	Co-worker Satisfaction	Allen, D. G., Bryant, P. C. and Vardaman, J. M. (2017)
		Cappelli, P., (2000)
		Chambers, E., G., Foulon, M., Handfield-Jones, H., Hankin, S. M., Michaels III, E.G. (1998)
	Persistence of low performers in the team	Chambers, E., G., Foulon, M., Handfield-Jones, H., Hankin, S. M., Michaels III, E.G. (1998)
	Team	Zhao, Y., Hryniewicki, M. K., Cheng, F., Fu, B., & Zhu, X. (2018)
	Team Loyalty	Cappelli, P., (2000)
		Farris, G. F. (1969)
		Rao, T.V. (2007)
	Team Size	Elvira and Cohen (2001)
		Baek, P. (2016)
		Farris, G. F. (1969)
	Competition between teams	Farris, G. F. (1969)
	Perception on co-worker's intention to leave	Kirschenbaum, A., Weisberg, J. (1990)
	Social community in workplace	Cappelli, P., (2000)
		Farris, G. F. (1969)
	Team based compensation	Cappelli, P., (2000)
	Perception of co-worker talent	Chambers, E., G., Foulon, M., Handfield-Jones, H., Hankin, S. M., Michaels III, E.G. (1998)
Stability	Company's performance	Chambers, E., G., Foulon, M., Handfield-Jones, H., Hankin, S. M., Michaels III, E.G. (1998)
	Company's market leadership	Chambers, E., G., Foulon, M., Handfield-Jones, H., Hankin, S. M., Michaels III, E.G. (1998)
	Company's job security perception	Chambers, E., G., Foulon, M., Handfield-Jones, H., Hankin, S. M., Michaels III, E.G. (1998)
	Company's prestige	Joseph, D., Ang, S., Slaughter, S. A. (2015).
		Ramesh, A., Gelfand, M. J. (2010).
	Closeness to retirement (public welfare)	Cappelli, P., (2000)
Rewards	Pay satisfaction	Allen, D. G., Bryant, P. C. and Vardaman, J. M. (2017)
	Recognition of contributions	Ajit, P. (2016)
		Rao, T.V. (2007)
	Salary and benefits	Ajit, P. (2016)
		Allen, D. G., Bryant, P. C. and Vardaman, J. M. (2017)
		Cappelli, P., (2000)
		Chambers, E., G., Foulon, M., Handfield-Jones, H., Hankin, S. M., Michaels III, E.G. (1998)
		Farris, G. F. (1969)
		Zhao, Y., Hryniewicki, M. K., Cheng, F., Fu, B., & Zhu, X. (2018)
	Time since last benefits increase	Cappelli, P., (2000)
		Zhao, Y., Hryniewicki, M. K., Cheng, F., Fu, B., & Zhu, X. (2018)
	Wage level	Kirschenbaum, A., Weisberg, J. (1990)
	Deffered compensation plan	Cappelli, P., (2000)
		Chambers, E., G., Foulon, M., Handfield-Jones, H., Hankin, S. M., Michaels III, E.G. (1998)
Life Quality	Closeness to kith and kin	Rao, T.V. (2007)
	Commute time	Rosenbaum, E. (2019)
	Location	Cappelli, P., (2000)
	Time of breaks	Baek, P. (2016)
	Geographic location	Chambers, E., G., Foulon, M., Handfield-Jones, H., Hankin, S. M., Michaels III, E.G. (1998)
	Perception of overload	Allen, D. G., Bryant, P. C. and Vardaman, J. M. (2017)
	Perception of respect for lifestyle	Chambers, E., G., Foulon, M., Handfield-Jones, H., Hankin, S. M., Michaels III, E.G. (1998)
	Perception of stress	Chambers, E., G., Foulon, M., Handfield-Jones, H., Hankin, S. M., Michaels III, E.G. (1998)
	Perception of work interference with family	Cappelli, P., (2000)
	Stress, anxiety, fatigue or burnout states	Ajit, P. (2016)
		Allen, D. G., Bryant, P. C. and Vardaman, J. M. (2017)

		Cappelli, P., (2000)
		Rao, T.V. (2007)
	Work hours and overtime	Farris, G. F. (1969)
		Huffman, A. H., Adler, A. B., Dolan, C. A., Castro, C. A. (2005).
		Rosenbaum, E. (2019)
Leadership	Leadership and Management	Ajit, P. (2016)
		Cappelli, P., (2000)
		Chambers, E., G., Foulon, M., Handfield-Jones, H., Hankin, S. M., Michaels III, E.G. (1998)
	Boss'Performance	Chambers, E., G., Foulon, M., Handfield-Jones, H., Hankin, S. M., Michaels III, E.G. (1998)
	Admiration for the Boss	Chambers, E., G., Foulon, M., Handfield-Jones, H., Hankin, S. M., Michaels III, E.G. (1998)
	Satisfaction with direct Supervisor	Allen, D. G., Bryant, P. C. and Vardaman, J. M. (2017)
		Cappelli, P., (2000)
		Rao, T.V. (2007)
	Satisfaction with communication with Leader	Allen, D. G., Bryant, P. C. and Vardaman, J. M. (2017)
	Direct Supervisor Turnover	Rao, T.V. (2007)
Job Characteristics	Accessory of business	Cappelli, P., (2000)
	Cross-disciplinarity of work groups	Baek, P. (2016)
		Farris, G. F. (1969)
	Department	Zhao, Y., Hryniewicki, M. K., Cheng, F., Fu, B., & Zhu, X. (2018)
	Independence or Autonomy	Allen, D. G., Bryant, P. C. and Vardaman, J. M. (2017)
		Cappelli, P., (2000)
		Chambers, E., G., Foulon, M., Handfield-Jones, H., Hankin, S. M., Michaels III, E.G. (1998)
		Farris, G. F. (1969)
		Rao, T.V. (2007)
	Is Client Facing Role	Zhao, Y., Hryniewicki, M. K., Cheng, F., Fu, B., & Zhu, X. (2018)
	Job Design	Allen, D. G., Bryant, P. C. and Vardaman, J. M. (2017)
		Cappelli, P., (2000)
		Rao, T.V. (2007)
	Job Satisfaction	Ajit, P. (2016)
		Allen, D. G., Bryant, P. C. and Vardaman, J. M. (2017)
		Jackofsky, E. F. (1984)
		Ramesh, A., Gelfand, M. J. (2010).
	Role Clarity	Spector, P. E. (1991).
		Allen, D. G., Bryant, P. C. and Vardaman, J. M. (2017)
		Rao, T.V. (2007)
	Role Conflict	Spector, P. E. (1991)
		Allen, D. G., Bryant, P. C. and Vardaman, J. M. (2017)
	Routinization or Repetitiveness	Parasuraman and Alutto (1984).
		Allen, D. G., Bryant, P. C. and Vardaman, J. M. (2017)
		Farris, G. F. (1969)
		Kirschenbaum, A., Weisberg, J. (1990)
	Sacrifice and dedication	Cappelli, P., (2000)
	Specialization/Complexity	Parasuraman and Alutto (1984).
		Farris, G. F. (1969)
		Jackofsky, E. F. (1984)
		Zhao, Y., Hryniewicki, M. K., Cheng, F., Fu, B., & Zhu, X. (2018)
	Title	Farris, G. F. (1969)
		Zhao, Y., Hryniewicki, M. K., Cheng, F., Fu, B., & Zhu, X. (2018)
	Interactions with other people (inside organization)	Baek, P. (2016)
		Farris, G. F. (1969)
	Interactions with other people (outside organization)	Farris, G. F. (1969)
	Time spent with people in meetings	Baek, P. (2016)
	Job perceived exciting challenges	Chambers, E., G., Foulon, M., Handfield-Jones, H., Hankin, S. M., Michaels III, E.G. (1998)
	Job Seniority Level	Cappelli, P., (2000)
Internal Growth Opportunities	Growth and self-fulfillment	Cappelli, P., (2000)
	Career stagnation	Rosenbaum, E. (2019)
	Promotion of peers	Rosenbaum, E. (2019)
	Long-range perspective	Farris, G. F. (1969)
	Perceived chances for improvement/development	Ajit, P. (2016)
		Chambers, E., G., Foulon, M., Handfield-Jones, H., Hankin, S. M., Michaels III, E.G. (1998)
		Farris, G. F. (1969)
		Kirschenbaum, A., Weisberg, J. (1990)
	Perceived exciting challenges in the company	Chambers, E., G., Foulon, M., Handfield-Jones, H., Hankin, S. M., Michaels III, E.G. (1998)
	Perceived chances for career growth	Allen, D. G., Bryant, P. C. and Vardaman, J. M. (2017)
		Chambers, E., G., Foulon, M., Handfield-Jones, H., Hankin, S. M., Michaels III, E.G. (1998)
Individual Performance	Performance	Chambers, E., G., Foulon, M., Handfield-Jones, H., Hankin, S. M., Michaels III, E.G. (1998)
		Farris, G. F. (1969)

		Jackofsky, E. F. (1984)
		Zhao, Y., Hryniewicki, M. K., Cheng, F., Fu, B., & Zhu, X. (2018)
Individual Characteristics	Sex	Ajit, P. (2016)
		Allen, D. G., Bryant, P. C. and Vardaman, J. M. (2017)
	Age	Ajit, P. (2016)
		Allen, D. G., Bryant, P. C. and Vardaman, J. M. (2017)
		Cappelli, P., (2000)
		Farris, G. F. (1969)
		Kirschenbaum, A., Weisberg, J. (1990)
		Zhao, Y., Hryniewicki, M. K., Cheng, F., Fu, B., & Zhu, X. (2018)
	Cognitive Ability	Allen, D. G., Bryant, P. C. and Vardaman, J. M. (2017)
	Management Level	Zhao, Y., Hryniewicki, M. K., Cheng, F., Fu, B., & Zhu, X. (2018)
	Marital Status	Ajit, P. (2016)
		Allen, D. G., Bryant, P. C. and Vardaman, J. M. (2017)
		Cappelli, P., (2000)
		Hom, P. W., Hulin, C. L. (1981)
	Personality factors	Cappelli, P., (2000)
		Rao, T.V. (2007)
	Race	Elvira and Cohen (2001)
		Ajit, P. (2016)
		Allen, D. G., Bryant, P. C. and Vardaman, J. M. (2017)
		Zhao, Y., Hryniewicki, M. K., Cheng, F., Fu, B., & Zhu, X. (2018)
	Tenure	Ajit, P. (2016)
		Allen, D. G., Bryant, P. C. and Vardaman, J. M. (2017)
		Cappelli, P., (2000)
		Kirschenbaum, A., Weisberg, J. (1990)
		Zhao, Y., Hryniewicki, M. K., Cheng, F., Fu, B., & Zhu, X. (2018)
	Children	Lee, T. W., Maurer, S. D. (1999).
		Blegen, M. A., Mueller, C. W., Price, J. L. (1988)
		Allen, D. G., Bryant, P. C. and Vardaman, J. M. (2017)
	Kinship Responsibilities (e.g. Small children, old parents)	Allen, D. G., Bryant, P. C. and Vardaman, J. M. (2017)
		Cappelli, P., (2000)
		Rao, T.V. (2007)
	Having a reference group outside the organization	Farris, G. F. (1969)
	Employee Seniority Level	Farris, G. F. (1969)
External conditions	Labor market conditions	Jackofsky, E. F. (1984)
	Perceived ease of movement	Allen, D. G., Bryant, P. C. and Vardaman, J. M. (2017)
		Farris, G. F. (1969)
		Kirschenbaum, A., Weisberg, J. (1990)
	Company reviews online	Baek, P. (2016)
Education & Training	Education	Bretz, R. D., Boudreau, J. W., Judge, T.A. (1994)
		Ajit, P. (2016)
		Allen, D. G., Bryant, P. C. and Vardaman, J. M. (2017)
		Farris, G. F. (1969)
		Zhao, Y., Hryniewicki, M. K., Cheng, F., Fu, B., & Zhu, X. (2018)
	ROI in Education	Rao, T.V. (2007)
	Training	Rao, T.V. (2007)
Corporate Culture, Values and Transparency	Communication and Fairness	Blegen, M. A., Mueller, C. W., Price, J. L. (1988)
		Erdogan, B., Bauer, T. N. (2010)
		Ajit, P. (2016)
		Allen, D. G., Bryant, P. C. and Vardaman, J. M. (2017)
		Cappelli, P., (2000)
		Farris, G. F. (1969)
	Realistic Job Previews	Allen, D. G., Bryant, P. C. and Vardaman, J. M. (2017)
	Identification with Mission, Values and Culture	Allen, D. G., Bryant, P. C. and Vardaman, J. M. (2017)
		Chambers, E., G., Foulon, M., Handfield-Jones, H., Hankin, S. M., Michaels III, E.G. (1998)
	Company loyalty	Farris, G. F. (1969)
	Recruitment Criteria	Allen, D. G., Bryant, P. C. and Vardaman, J. M. (2017)
	Company satisfaction	Farris, G. F. (1969)