ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ

ΤΜΗΜΑ ΔΙΟΙΚΗΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

ΘΕΟΦΙΛΑΤΟΥ ΖΩΗΣ

ΤΙΤΛΟΣ ΔΙΠΛΩΜΑΤΙΚΗΣ

SURVIVAL ANALYSIS & MACHINE LEARNING ON EMPLOYEE TURNOVER

Επιβλέπων :
Νικόλαος Κορφιάτης
Επισκέπτης Αναπληρωτής Καθηγητής
Επιχειρηματικής Αναλυτικής

Υποβληθείσα ως μέρος των απαιτήσεων για την απόκτηση
Μεταπτυχιακού Διπλώματος (MSc) στη Διοικητική Επιστήμη και Τεχνολογία

Αθήνα, Μάρτιος 2022


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Βεβαίωση εκπόνησης Διπλωματικής εργασίας

«Δηλώνω υπεύθυνα ότι η συγκεκριμένη μεταπτυχιακή εργασία για τη λήψη του μεταπτυχιακού τίτλου σπουδών του ΠΜΣ στη Διοικητική Επιστήμη και Τεχνολογία του Τμήματος Διοικητικής Επιστήμης και Τεχνολογίας του Οικονομικού Πανεπιστημίου Αθηνών έχει συγγραφεί από εμένα προσωπικά και δεν έχει υποβληθεί ούτε έχει εγκριθεί στο πλαίσιο κάποιου άλλου μεταπτυχιακού ή προπτυχιακού τίτλου σπουδών στην Ελλάδα ή το εξωτερικό. Η εργασία αυτή έχοντας εκπονηθεί από εμένα, αντιπροσωπεύει τις προσωπικές μου απόψεις επί του θέματος. Οι πηγές στις οποίες ανέτρεξα για την εκπόνηση της συγκεκριμένης διπλωματικής αναφέρονται στο σύνολό τους, δίνοντας πλήρεις αναφορές στους συγγραφείς, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το διαδίκτυο».

(Υπογραφή)

ΘΕΟΦΙΛΑΤΟΥ ΖΩΗ

Φοιτητήτρια MSc στη Διοικητική Επιστήμη και Τεχνολογία

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Abstract

Employee attrition may be genuine concern in information-based organizations. When representatives take off an organization, they carry essential inferred information that is regularly the source of competitive advantage for the firms. For an organization to persistently have the next competitive advantage over this competition, it ought to make it an obligation to minimize employee steady loss. Numerous components might be responsible for that (for case: social, budgetary, ramp downs, managerial issues etc.). Each organization has a certain way to behave its employees and assure their satisfaction with all aspects. However, frequently no measures are taken as regards the fulfillment rate. Therefore, in numerous cases, personnel resigned suddenly without an apparent justification. This proposal investigates the effectiveness of survival analysis & machine learning algorithms in forecasting employee attrition. In specific, the current thesis mentions the predictive performance of the Cox Proportional Hazards - Model and some machine learning algorithms such as DeepSurv, Random Survival Forest, max -out. To illustrate the impact of the above to employee turnover, IBM's synthetic workforce data from Kaggle were used where the Random Forest algorithm was applied to predict the future turnover of employees.

Keywords: Employee attrition, Cox Proportional Hazards - Model, Random Survival Forest

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Table of contents

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

1

Introduction

1.1 Voluntary turnover nowadays

Voluntary turnover and employee attrition presents a major issue for companies as it influences not just their operations and work manageability yet additionally their drawn-out development systems. On this path, employee retention constitutes a major challenge for recruitment departments of each company, since organization's attrition implies not only the deficiency of abilities, encounters, and workforce, but the loss of business opportunities due to lack of resources. In the era of machine learning and big data, *people analytics* help organizations , as well as human resources supervisors, to reduce attrition by changing the approach or attracting and retaining talent to a data-driven and evidence based one. Predicting employee churn, and understanding its driving variables via data analytics, might help both the organizations and the HR personnel. For instance, since the reason of a resignation may be prolonged work hours, overtime occurred unexpectedly the business may proactively take actions related with shift scheduling and retain employees who may be at risk of quitting and save on turnover costs.

1.2 Research Objective

Given the current context, this thesis aims at anticipating voluntary employee attrition and explores the transcendent reasons that generate employee turnover. To do this, it will utilize *survival analysis* which is a branch of statistics for analyzing the anticipated length of time until an event occurs. Survival analysis requires special techniques since there is the possibility of not observing an event of interest for some individuals, this is commonly known as truncated data. For instance, some individuals may have a different event or have to drop out of the study. To handle these incomplete observations, that cannot be ignored survival analysis uses censoring. In the case of employee turnover, right censored data are present: *if an employee has*

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

not resigned at the time of the data acquisition this doesn't guarantee that he will not leave in the future. It is worth emphasizing that machine learning does not deal with censoring; for this reason study has focused both on survival analysis and ML methods in order to be conducted.

1.2.1 Contribution and intended audience.

This research encloses a methodology to predict employee voluntary employee attrition. By all means, numerous machine learning algorithms have been extended in order to deal with censored data and perform survival analysis. Nevertheless, novel techniques have not be used so that they can predict voluntary employee turnover so far. In this thesis we will mention the contribution of Cox Proportional Hazard Model (Cox PH) and compare the last with machine learning algorithms such as Random Survival Forest and max-out. Therefore, we will utilize the best performing survival model to investigate the reasons employees leave on an association, hence empower in that way the business to act proactively and produce the employer's culture advantages that will result to higher employees' retention rate.

To predict voluntary employee attrition and search through it is main driving causes, we are going to use the IBM's data set provided in Kaggle called "*IBM HR Analytics Employee Attrition & Performance*". The inspiration behind using this data set with regards to Survival Analysis it is that the last includes both an occasion (Attrition) and a respective time period variable (YearsAtCompany). Certainly, both are required for a conduction of survival analysis. Moreover, we selected this data frame since the number of covariates it holds is appropriate for getting meaningful insights in regard to the reasons of conscious employee churn.

The contribution of this thesis is summarized as follows:

1. We evaluated the performance of Random Survival Forest in IBM's HR dataset
2. We studied two different approaches; survival analysis & max - out ML algorithm.
3. We evaluated the performance of each algorithm.

1.3 Thesis Structure

In this part we describe the chapters of this thesis.

More specifically:

- The relevand literature is discussed in Chapter 2.
- Chapter 3 discusses the main turnover meanings and models as well as definitions of survival & machine learning algorithms that will be implemented through our research to predict the voluntary employee turnover.
- Chapter 4 develops our experimental set up where Random Forest Classifier is implemented for the prediction of the attrition rate in IBM's data set.
- Chapter 5 explains in detail how similar experiments were processed using two different approaches.
- Chapter 6 gives the evaluation of both methods; Machine Learning Algorithms & Survival Analysis for predicting employee attrition and provides thoughts for further research.
- Chapter 7 provides details regarding the simulation environment required and source of the dataset used for the experimental set up processed through Chapter 4.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

2

Related work

We will focus on two domain thematic areas in relation to previous research that has been conducted and is relevant to the topic of this thesis. The first one concerns the use of machine learning models for predicting employee attrition in an organization and the second one concerns the use of statistical methods to accomplish this scope. In this thesis we used both aforementioned methods and compare them to extract the best possible results.

2.1 Machine Learning Algorithms to predict attrition rate

Previous research using machine learning algorithms has been made in order to predict employee attrition rate. For instance, Alao D. & Ademeyo A.B (2013) have conducted the research "*Analyzing Employee Attrition using Decision Tree algorithms*". Through this study the last used an employee dataset which included three hundred and nine (309) complete records of employees of one of the Higher Institutions in Nigeria who were employed and left between 1978 and 2006. In their method they used Decision Tree learning, which is a method commonly used in data mining, also the algorithm Pseudocode that is the general algorithm for building decision trees and eventually for the development of the relevant employee prediction model they used machine learning software such as WEKA written in Java. Results obtained from their study showed that employee salary and length of service where the main driving factors for predicting employee attrition.

Another research conducted and is related to this thesis, was the one of Ricardo Colomo – Palacios, Nesrine Ben Yahia & Jihen Hlel (April 20,2021) named "From Big Data to Deep Data to Support People Analytics for Employee Attrition Prediction". Thereby in their paper they aimed to propose a deep data driven predictive approach that could predict employee intention to leave an organization. They used also, HR IBM and HR Kaggle datasets in order to conduct their research. The methods used where Machine, Ensemble and Deep learning

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

techniques (ML, EL and DL) so that they were able to make interpretations and analyze further the exact root cause behind employee churn.

Additional research previously made using machine learning techniques was the one of M. Ravi (February 2020) with the subject "Prediction of Employee Attrition using Random Forest Classifier Technique". In this thesis, he used some methodologies of data classification. Those methodologies were Decision Tree and Naive Bayes (a classification technique depending on Bayes Theorem). For his project he also used the data from IBM containing 1470 records and 35 field of categories. The results of his thesis described that data extraction algorithms can be used to construct reliable methods for employee turnover.

2.2 Statistical models to predict employee churn

Researchers also applied methods other than machine learning algorithms such as statistical models. For instance, Paula C. Morrow, James C. McElroy, James B.Fenton & Kathleen S. Laczniak in their study named " Using Abesenteeism and Performance to Predict Employee Turnover: Early Detection through Company Records" (1999) demonstrated that absenteeism as measured by sickness absences relates positively to voluntary turnover and that performance relates negatively to the resignation rates of employees within an organization. These findings were made based on logistic regression; a statistical technique uses a logic function to model a binary dependent variable (Wikipedia, Logistic regression).

Another approach of research that has been conducted in regard to the prediction of voluntary employee churn contains survival analysis. The last consists a collection of statistical procedures for data analysis where the outcome variable of interest is time until the event occurs (Survival Analysis Part I: Basic concepts and first analyses, Br J Cancer. 2003 Jul 21; 89(2): 232–238, Published online 2003 Jul 15.). For instance, the study conducted from Christopher E. Penney (October 2016) with the subject "A survival analysis of ADM (Materiel) workforce attrition. This paper explored the problem of employee attrition and the role of financial factors in attrition behavior. The method used was survival analysis and in particular the Kaplan – Meier Model, the Cox Proportional Hazards Model & the Extended Cox Proportional Hazards Model.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

3

Theoretical Background

In this research, we chose to utilize Random Forest algorithm to display voluntary employee churn. We also wanted to highlight the importance of the survival analysis and its contribution to the prediction of voluntary employee churn. Specifically, we decided to analyze among others, survival analysis techniques as they offer the ability to manage censor data, though other methods, such as machine learning algorithms don't. In a dataset, censoring refers to the incomplete observation of an event being studied (Moore, 2016). In particular, in cases like the voluntary employee churn right censoring is present. In order to include any censor data in our estimates and succeed a more accurate approach we preferred to describe in detail survival methods over other approaches or integrated with ML algorithms. Within the current section, we present the main turnover meanings and models as well as definitions of survival & machine learning algorithms that will be implemented through our research to predict the voluntary employee turnover.

3.1 Defining Turnover

3.1.1 Staff turnover

The issue of staff turnover is at the forefront of the organizational behavior literature. More than a hundred papers have been published in prominent management journals and subject areas over the past decade, suggesting growing interest in the topic.

3.1.2 Definitions and forms of staff turnover

The phrase turnover is often used to characterize the departure of an employee from an organization Price 2001. It is usually described as the number of staff joining and leaving a firm.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

in a given period of time. Voluntary and involuntary staff turnover are the two types of staff turnover. In the first case, an employee voluntarily leaves the organization for reasons that vary from employee to employee.

Involuntary turnover occurs when a firm decides to fire an employee for a specific reason (financial, technological, poor performance or serious misconduct.). In other words, involuntary turnover occurs as a result of layoffs, while voluntary turnover occurs when employees choose to leave the organization on their own. The importance of this organizational phenomenon is such that there is a lot of research on the subject. Most of these studies (Wright, 1993) have focused on voluntary versus involuntary turnover. As a result, we can conclude that voluntary turnover is a serious issue for both individuals and firms. The voluntary turnover is the total number of employees leaving the firm divided by the total number of employees, usually over a one-year period (Hausknecht & Trevor, 2017).

Given the magnitude of the voluntary turnover problem for firms, academics and practitioners have paid particular attention to it in their study and organizational strategies. According to Armstong (2012), turnover can be a disruptive event with significant implications for the firm. This assumption is supported by the fact that an employee's decision to leave voluntarily is more likely to have negative consequences for the organization that firing a low performing employee.

3.1.3 Voluntary turnover costs

The issues of personnel turnover have been continual challenge for human resources managers and companies in whatever economy. A qualified and experienced employee's resignation is a very costly act for the company (Shaw and Dess, 2001).

Voluntary departure has a variety of consequences, including financial costs that affect the financial situation of the organization and moral costs that can affect the well-being and productivity of employees:

- Overtime has psychological effects on employees who remain in the organization. This could lead to tensions which in turn could have an impact on the social atmosphere (Grissom, et al., 2012; Wang, et al., 2012).
- The loss of an employee means the loss of a productive member that can destabilize the work environment.
- The quality of services offered to consumers is being decreased
- The risk of hiring and the risk of hiring an undervalued or poorly integrated new employee, as well as the cost of training new employees.
- The resigning employee's assets, knowledge, and accomplishments are tough to regain for the company. As a result, the new employee's learning curve lengthens.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Voluntary turnover on the other hand, does not have to be viewed as a negative thing. In truth, the exiting employee's turnover allows the company to renew and rejuvenate its team on the one hand, while allowing the outgoing employee to choose another job that best suits his abilities and aspirations on the other.

3.1.4 Process of voluntary turnover

Mobley (1978) presents diagrammatically a set of forms that make up the decision-making process in preparation for take-off. An assessment stage of the existing business gives rise to the strategy. An employee investigates his or her performance based on the underwriting criteria after a series of vital elements. These enable him or her to create the behavior he or she should take towards his or her job. The employee's ability to remain his current position is unexpected for his job bliss. The proximity of this gratification guarantees that the employee will remain in his current job. On the other hand, when a co-worker becomes frustrated, he reacts by separating himself from his responsibilities. This does so by taking inactive behaviors at work, which makes him less competent. At this level the employee is complimented by the thought of stopping his work, particularly in case the feeling of disappointment continues.

At a certain point, Michaels, and Spector (1982) included an additional variable: organizational commitment. Both studies reveal an employee who does not have a strong attachment to his or her organization begins to weigh the costs and benefits of seeking a new job. As a result, when the employee concludes from the assessment that he or she has a good chance of finding other employment, he or she explores for new opportunities outside his or her company. Once this stage of exploration is completed, he or she carries out an evaluation and comparison of the available options.

In this way, the partner will create an intention to either stay or leave the position or indeed be an organization. In the event that he or she has developed the intention to leave, the final stage will be to start. This demonstration makes it clear that the intention to leave the organization can emerge outside of decision-making preparation. In this case, the representative makes the choice in a sudden and highly tactically imprudent manner. This leads him to resign immediately instead of taking time to rationalize his choice to leave. Therefore, frustration in a method organization can bring a person back to a previous organization.

3.1.5 Intention of voluntary turnover

Mobley's performance emphasizes the intrigue that can be considered and the incorporation of deliberate flight as a nodal and rapid vector of successful flight, turning to the assumption of thoughtful activity. First presented and created by Fishbein (1967). This theory, which seeks to

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

decide the relationship between an individual's intentional behavior and the actual achievement of that behavior, is widely used by numerous analysts in some fields.

The theory of reasoned action explains how to understand a person's voluntary behavior and argues that the best approach to predicting voluntary behavior is through intention, which is a cognitive representation of a person's desire to do an activity. It is thought to be the immediate precursor to the realization of the behavior. Fishbein is a character in the movie Fishbein (1967). Indeed, the behavioral goal is grounded in the subjective rules and attitudes about the activity. The individual's favorable or unfavorable evaluation of the performance of the intended activity is referred to as behavioral attitude. It is determined by a review of the individual's views on the consequences and possibilities of achieving a particular behavior.

The subjective rule relates to a person's perception of whether to perform a particular behavior. According to this hypothesis, the more important people in a person's life believe that they should or should not perform an action, the more likely they are to do or not to do it. Individuals' predisposition to engage in an activity, however, does not always lead to the realization or concretization of their intention to reality. This raises the question of whether it is necessary to introduce the degree or behavioral control.

Indeed, Ajzen has extended the theory of reasoned action to the theory of planned action by integrating perceived behavioral control, which is defined as the existence or absence of elements that may promote or hinder behavior achievement.

In other words, behavioral control beliefs refer to a person's judgments about his or her ability to perform a specific conduct as a factor in behavioral intent. Thus, in an organizational environment, the emergence of the intention of voluntary turnover happens when a collaborator seriously and purposefully considers leaving his organization. Indeed, the goal of voluntary turnover has piqued the curiosity of scholars. This interest on the part of academics can be explained by the fact that various studies have established the importance of this desire as an immediate factor of voluntary turnover.

As a result, the intent of the roll, according to the planned action theory, becomes successful when individuals believe they have control over the decision to quit. In the same vein, there are other reasons why people may believe they have less influence over their decision. Griffeth and Peter, 2004, describe that:

- The decision of leave may be hampered by family or financial limitations
- Individuals are increasingly investing in a company over time, making it more difficult to depart (Becker 1960)
- Control perceptions may be influenced by perceptions of alternative availability and quality.

3.2 Turnover Models

3.2.1 March And Simon's Model

Overview: March and Simon's (1958) framework are used in many studies of voluntary attrition. These models can be traced back to Barnard-Simon's theory where they argued that all workers are faced with decisions through their interactions with their company (Mano, 1994). So 'key participation decision variable such as 'desire and ease of movement in and out of the organization is of particular interest in this study (Bowel and Siehl, 1997). Given this decision-making ability, this theory further identifies this employee's decision to quit is caused by two factors such as perceived ease of movement and perceived desire to move. However, perceived ease of movement refers to the evaluation of perceived alternatives or opportunities while perceived desire to move is influenced, for example, by job satisfaction (Morell et al. 2001, Samad and Yusuf, 2012). Nevertheless, it was found that when incentives are increased by the company, this will decrease the tendency of employees to leave and vice versa (Morell et al. 2001). In this model, the possibility internal turnover is considered in advance before external turnover is decided.

Limitations and recommendations: According to the authors some limitations prevent them from recommending March and Simon's model. When describing the turnover process, the term "turnover" is used. The model merely shows a static perspective of the situation rather than a process perspective of the attrition. They also omit critical factors that influence turnover, such as a role stress or job satisfaction and organizational commitment in various forms (Morell et al. 2001; Allen and Shannock, 2012). In addition, factors related to employee turnover, such as according to some hypotheses, March and Simon's model has had a significant impact on subsequent research on the subject. Other components of the study may be hampered by staff turnover and success. As a result, we are unable to understand the relationship between this model and organizational commitment and how it affects the turnover process.

3.2.2 Porter and Steers (1973) Met Expectations Model

Overview: The expectations were fulfilled by Porters-Steers. Vroom's theory of expectations was modified into a hypothesis. Porters and Steers identified three common denominators in motivation. The three numerators are as follows: (a) what triggers human behavior; (b) what directs or channels that behavior; and (c) how that behavior is maintained and in the long term. Needs or expectations, they believe, are the primary building blocks of a motivational model.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Goals, behavior and some kind of feedback. The term 'discrepancy' is used to describe the concept of satisfied expectations between what a person, experiences at work, either positively or negatively, and what they expect to encounter (Porter and Steers, 1973). When a person's expectations – whatever they may be – are met, this assumption holds. If his needs are not met, he is more likely to withdraw. They also claimed that the level of job satisfaction was a factor. They reported the total number of employee expectations that were met. Motivation theory has been studied for many years and there are various theories and definitions of motivation. Vroom's expectancy, according to Vroom, is an "instantaneous belief about the likelihood that a particular action will take place followed by a particular outcome". Thus, Vroom's expectancy theory can be used to predict the degree of work job satisfaction (Samad and Yusuf, 2012).

Limitations and Recommendations: Porter and Steer's Met Expectations Hypothesis focused on a single antecedent of turnover. Potential moderating effects on turnover decisions are not identified. Thus, instead of focusing merely on internal factors, a turnover model that considers various antecedents should be considered.

3.2.3 Price (2001) Causal Model of Turnover

Authors such as James Price, Charles Muller and others have constructed models that identify the determinants of voluntary and work groups, having conducted 33 studies over the past four decades. The causal model of employee voluntary turnout (Price, 2001) is a brief reflection on the factors that influence employee voluntary turnout. Price divided the causal factors into external variables and intrinsic intervening variables. The exogenous variables in the model are further classified into environmental, individual and structural labels. Intervening variables, on the other hand, are considered endogenous variables. Non – work environments impose limits on the intention to stay, which are represented by the environment variables. The availability of alternative occupations in a career environment is referred to as opportunity (Price, 2001; Boyar et al., 2012). It has been discovered to have a direct effect. There is a positive correlation to the intention to move. To put it another way, more opportunities lead to more turnover. Workers will be more aware of alternative occupations that are accessible in their industries as a result of this objective. After those employees will analyze the costs, risks, and rewards of different career options. When the other occupation offers a higher pay package, employees may be more dissatisfied with their current position as a result, which could lead them to resign. The second environmental element is kinship responsibility, which refers to responsibilities towards family members, residing within the community (Boyar et al., 2012). It has been discovered that kinship obligation has a direct negative relationship with turnover. When a family lives close to an employee's workplace, this can create a sense of obligation in the individual, which can be easily satisfied by continuing to work in their current job. As a result,

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

the likelihood of staff leaving is reduced (Hom and Griffeth, 1991; Allen and Shannock 2012; Samad and Yusuf, 2012). General training, job involvement, positive emotionality and negative emotionality are four separate extrinsic characteristics identified by Price as directly influencing turnover intention. Turnover probability is significantly influenced by general training. Increased general training, according to Price and Mueller, leads to a higher turnover rate.

Limitations and recommendations

However, the turnover model developed by Price (2001) has some weaknesses. This model, for example, ignored the process and effect of intervening and moderating variables, the narrowness and homogeneity of the study populations, the failure to detect differences in behavior between part time and full-time employees, and the lack of longitudinal research on the data collected (Goodman, 2007). Consequently, in addition to the exogenous and endogenous variables contained in Price's turnover model, a more predictive turnover model should include the criteria described above.

3.2.4 Mobley (1977) Intermediate Linkages Model

Overview: Mobley (1977) was the first to present a comprehensive diagram of the process of psychological change. March and Simon's theory of job ease and desirability, and Porter and Steer's model of expectancy satisfaction and desire to leave, are based on Mobley's model. Rather than being descriptive, this model is heuristic.

According to Mobley's theoretical model, the primary process for converting dissatisfaction into actual turnover is based on three turnover insights:

1. Consider leaving – An employee is considering leaving the company.
2. Intention to seek – An employee decides to seek employment outside the company.
3. Intention to resign – An employee has made the decision to resign from the company at some uncertain point in the future.

Withdrawal behavior, according to Mobley's theory, is a branch of decision making associated with a series of cognitive stages that begin with an evaluation of the current task and culminate in an emotional state of satisfaction or dissatisfaction (Mobley et al.,1979; Thwala et al., 2012; Abdullah et al.,2011). Only then will your dissatisfaction prompt you to consider resigning. Only after searching for alternatives, evaluating these alternatives and comparing them with the current job can employees create the intention to leave. Finally, the individual will either leave or stay.

Limitations and Recommendations: The limitation of the Mobley model is that it is intended to be used only as a starting point for the construction of later models. Mobley's (1977) model included only the links of the turnover process from individual intentions to the present and did

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

not include elements of turnover from aspects of job satisfaction or organizational commitment factors (Allen and Shannock, 2012; Samad and Yusuf, 2012). However, in Price and Mueller, the following improvements are made to the model to further extend Mobley's model by incorporating other variables such as organizational commitment and other investigated aspects that more specifically affect job satisfaction.

3.2.5 Sheridan and Abelson (1983) Cusp Catastrophe Model of Turnover

The expected correlations between job satisfaction, job strain, organizational commitment and actual employee turnover are linearized in this model (Holtom et al., 2006; Mitchel and Lee, 2001; Messersmith, 2007). There are three primary properties about it. First this model, portrays withdrawal behavior as a discontinuous and dynamic process caused by rapid changes, which differs from Mobley's turnover process in a surprising way. However, it is also mentioned that this model uses the delay rule, which states that employees would try to stay with the company for as long as possible. Rather than following the intermediate ties model, the Cusp model believes that departure is governed by two factors: a disruptive factor and an attractor. Job strain is an example of a disruptive factor, while the term attractor refers to the loyalty of an organization.

Occupational tension is divided into three parts: role conflict, role ambiguity and role precision, according to a more detailed description of the separation factor. The disagreement and gap between external and internal roles can be called role conflict between and internal customer and an external customer, as well as the objects and services that the employee can provide. The uncertainty about the role signals that the employee is not sure how the job should be done or how a person should be evaluated. Rewards are given. Finally, role correctness refers to employees' perceptions of what their bosses expect that flow from them (Thurau, 2000). Organizational commitment involves team cohesion, which is magnet here as well as friendliness and cooperation among colleagues (Thurau, 2000). Emotional commitment is a term used to describe positive organization commitment (Allen and Shannock 2012; Samad and Yusuf 2012). The factors that influence organizational commitment should then be based on the factors that influence job satisfaction in an organization, which means that when continuous commitment becomes fiscal, turnover intention will result (George and Jones, 2012).

Limitations and Recommendations: The model's weaknesses include that it assumes that organizational behavior is too qualitative, and that predicting the behavior of even the most basic complex organizations remains a challenging task.

Furthermore, this model is very complex and fails to adequately describe a complex system with numerous critical variables. It is extremely improbable to be able to forecast every aspect of the behavior of extremely complex organizational structures. Cusp Catastrophe Model could

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

be condensed to a simpler version that is easily understanding only by looking at the recommended variables.

3.2.6 An Integrated Process Model (Jackofsky, 1984)

Overview: The Integrated Process Model was one of Ellen F. Jackofsky's turnover models and some of the key elements of this model can be traced back to March and Simon's (1958) model published by Ellen F. Jackofsky. It consisted of desirability and mobility. Ease of mobility was cited as the basic model. The perspectives and alternatives of other organizations, as well as extra-organizational elements such as good work, are all elements to be considered. Market conditions, employee's satisfaction, and desired mobility on the other hand, are inferred. For the most part, this referred to possible intra – organizational issues. These two developments were viewed as major causes that might encourage employees to voluntarily leave the organization. However, there were significant flaws in this fundamental model, as other crucial elements, such as job performance, were not properly examined, which accelerated the turnover process. As a result, additional studies were conducted by incorporating the impacts of job performance into the fundamental voluntary turnover model with several hypotheses that resulted in involuntary and voluntary turnover.

Organizational characteristics include the importance of motivation, the structure of assigned tasks and the leader's behavior, according to Jackofsky (1984), while personal characteristics include individual level of competence and self – esteem, commonly known as individual differences. Without a doubt, the importance of motivation has been linked to a job performance (Solomon et.al 2012; Abdullah et.al. 2011). In general, competent employees would be rewarded with larger bonuses, year-end travel vouchers and other incentives than those who performed poorly. These benefits would undoubtedly increase their contribution to their organization; however, this aspect, is likely to be negatively related or unrelated to job performance, as strong employees may be offered better employment opportunities elsewhere and leave to work for a competitor.

Limitations and Recommendations: It is clear that the relationship between job performance and employee's intention to move is critical in determining the impact of employee's resignation, whether the relationship is good or negative. In this model, potential determinants and job performance are projected to be closely intertwined. However, a turnover intention survey based on job performance and other job determinants is not sufficient to predict an organization's turnover intention without taking into account additional elements such as individual and extra-organizational factors.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

3.2.7 Mitchell & Lee (2001) – Job Embeddedness Model

Overview: The turnover unfolding model proposed by Lee and Mitchell (1994) is based on Beach's image theory, a 1990s theory of general decision making. Individuals sort and interpret information as they evaluate options, according to image theory. As individuals quickly sort through information, they determine the degree to which the information is accurate.

The information corresponds to employee's perceptions of value, course, and tactics. The sorting is carried out only on the basis of the term "fit violation" refers to the determination of how well the environment fits an individual's personality. The process usually begins with a separate event. The incident prompts a person to stop and consider the significance or implications of the incident in relation to their work. As a result, when individuals believe that leaving is an option worth considering, they will explore whether there are other options. As previously mentioned, the turnover process begins with individuals detecting significant incidents, as Lee and Mitchell (1994) concluded. Individuals nowadays are constantly asking themselves whether employment offers the financial rewards or rewards they desire. If their demands are not met, they become dissatisfied with their jobs, which prompts them to seek and analyze opportunities (Mitchell and Lee, 2001; Holtom et al.2006; Allen and Shannock, 2012). Even if a person has looked for another job or has a job offer, he or she must decide whether to leave voluntarily or stay in the current job. Furthermore, these psychological processes revealed that individuals often adopt one of five hypothetical decision options and execute their response. The first path reflects those who are shocked but have no plan of action or are seeking alternatives. The second pathway occurs when people are shocked and this is seen as a violation of their belief or images. People who follow the third path are more conscious in their decision to depart.

Limitations and Recommendations: In terms of theoretical hypotheses and empirical research, the model of (Lee and Mitchell, 1994; Mitchell and Lee, 2001) suggests that job coupling plays a role similar to job satisfaction and organizational commitment and may act as a more decisive mediating variable that directly leads to voluntary employee turnover under certain conditions. On the other hand, traditional attitude models ignore the importance of this type of influencing factors. In addition to this, a multi-path model for the work-co-work link could provide further benefits for assessing actual turnover behavior and could be useful in extending the scope of organizational behavior. Job-job coupling is a new variable included in the standard model that advocates in two dimensions: on-job coupling and off-job coupling. Connection, appropriateness and sacrifice are the three primary structural variables of labor coupling. Combining the social background from the perspective of turnover decisions with multi-path analysis, Lee and Mitchell (1994) proposed and refined the "unwrapped job-coupling model" of voluntary turnover for worker retention. The researchers found that job coupling has a greater

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

impact on employee retention or voluntary turnover than job satisfaction or organizational commitment (Mitchell and Lee, 2001).

Allen and Griffeth (2003) divided their findings into three categories based on their research on the effect of employee performance levels on voluntary turnover. These were developed from classical organizational equilibrium theory (March and Simon, 1958) and media chain process theory of turnover and then presented a comparatively complete unifying research model for discussing the relationship between the employee performance and the tendency to withdraw and voluntary turnover (Messersmith, 2007; Hom and Griffeth, 1991). As a result, the proposed model consists of three analytical pathways: The performance character of employees in organizations will affect their job satisfaction and organizational commitment in two ways. First, the performance character of employees' performance in organizations will influence their labor market mobility behavior through labor market mobility with perceived ease of mobility as a determinant variable; second, the performance character of employees' performance in organizations will influence their labor market mobility with perceived ease of mobility as a determinant variable; the third theme is 'short – circuiting' which refers to the way in which different levels of employee performance in organizations affect turnover behavior in a more direct way (Mobley et al.,1979, Lee and Mitchell, 1994). Compared to previous models, this is a little better.

3.3 Survival & Machine Learning Algorithms

Prior mentioning the algorithms used in this research, we specify some survival analysis definitions useful for their description. The first of these is the survival function. The survival function simply indicates the probability that the event of interest has not yet occurred by time t; thus, if T denotes time until resignation, S(t) denotes probability of staying within the organization beyond time t (BIO 244: Unit 1 Survival Distributions, Hazard Functions, Cumulative Hazards). It is denoted as follows:

$$S(t) = 1 - F(t) = P\,(T > t) \text{ for } t > 0 \quad (1)$$

Another one useful term is the hazard function; denoted as λ or h, is defined as the event rate at time t conditional on survival until time t or later (that is, T >= t). Suppose that an employee hasn't left the company for a time t and we desire the probability that he will leave for an additional time dt (https://en.wikipedia.org/wiki/Survival_analysis) :

$$\lambda\,(t) = \lim_{dt \to 0} \frac{\Pr(t \le T < t + dt)}{dt * S(t)} = \frac{f(t)}{s(t)} = -\frac{S\prime(t)}{S(t)} \quad (2)$$

The hazard function can alternatively be represented in terms of the cumulative hazard function, conventionally denoted Λ or H (https://en.wikipedia.org/wiki/Survival_analysis):

$$\Lambda(t) = -\log S(t)$$

so transposing signs and exponentiating

$$S(t) = \exp(-\Lambda(t))$$

or differentiating

$$\frac{d}{dt}\Lambda(t) = -\frac{S\prime(t)}{S(t)} = \lambda(t)$$

The name "cumulative hazard function" is derived from the fact that

$$\Lambda(t) = \int_0^t \lambda(u)du$$

which is the "accumulation" of the hazard over time.

3.3.1 Cox PH

The cox (proportional hazards or PH) model (Cox, 1972) it is a survival survival regression model, which depicts the connection between the occasion incidence, as communicated by hazard function and a set of covariates.

Mathematically the Cox model is denoted as:

$$h(t) = h_0(t)\ x \exp\{b_1x_1 + b_2x_2 + \ldots + b_px_p\}$$

where the hazard function h(t) is dependent on (or determined by) a set of p covariates $(x_1, x_2, \ldots, x_p)$, whose impact is measured by the size of the respective coefficients $(b_1, b_2, \ldots, b_p)$. The term h_0 is called the baseline hazard, and is the value of hazard if all the x_i are equal to zero (the quantity exp(0) equals 1). The 't' in h(t) reminds us that the hazard may (and probably will) vary over time. An appealing feature of the Cox model is that the baseline hazard function is estimated nonparametrically, and so unlike most other statistical models, the survival times are not assumed to follow a particular statistical distribution *(Survival Analysis Part II: Multivariate data analysis – an introduction to concepts and methods, British Journal of Cancer 2003, MJ Bradburn*,1, TG Clark1, SB Love1 and DG Altman1)*.

The Cox model is basically a multiple linear regression of the logarithm of the hazard on the variables x_i with the baseline hazard being an 'intercept' term that varies with time. The covariates at that point act multiplicatively on the hazard at any point in time, and this provides

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

us the key presumption of the PH model: the hazard of the event in any group is a constant multiple of the hazard in any other *(Survival Analysis Part II: Multivariate data analysis – an introduction to concepts and methods, British Journal of Cancer 2003, MJ Bradburn*,1, TG Clark1, SB Love1 and DG Altman1)*.

3.3.2 DeepSurv

DeepSurv consists of a Cox proportional hazards deep neural network which predicts the effects of an employee's covariates on their hazard rate parameterized by the weights of the network θ. It is denoted as follows:

$$l(\theta) = -\frac{1}{N_{E=1}} \sum_{i:E_i=1} (\hat{h}_\theta(x_i) - \log \sum_{j \in R(Ti)} e^{\hat{h}_\theta}(x_j) + \lambda \times || \theta ||_2^2$$

where the $N_{E=1}$ is the number of employees with an observable event and λ is the l2 regularization parameter (Katzman et al. BMC Medical Research Methodology 2018)

3.3.3 Random Forest Algorithm

Random forest algorithm is a machine learning technique developed by Lao Breiman and Adele Cutler that combines the output of numerous decision trees to produce a single outcome. Its popularity is due to its ease of use and adaptability since it can handle both classification and regression problems.

3.3.3.1 Decision trees

As the random forest model is made up of several decision trees, it is a good idea to start with a brief description of the decision tree algorithm. "Should I surf" for example, is a good starting point for a decision tree. To get an answer you can ask series of questions such as "Is it a long period as well". These questions serve as a decision node in the tree, allowing the data to be separated. Each inquiry aids a person in reaching a final conclusion, which is indicated by the leaf node. Observations that meet the requirements will be routed down the "No" branch. The goal of the decision trees is to find a solution (source: https://www.ibm.com/cloud/learn/random-forest).

Decision trees are examples of classification problems. While they are ubiquitous supervised learning methods, they may suffer from problems of bias and overfitting. The random forest technique on the other hand, predicts more accurate results when multiple decisions trees form a set especially when individual trees are uncorrelated.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

3.3.3.2 Ensemble methods

Ensemble learning strategies consist of a set of classifiers – e.g., decision trees – and their predictions are summed to discriminate the most dominant outcome. The best-known collection strategies are sacking, also known as bootstrap clustering, and boosting. In 1996, Leo Breiman introduced the bagging strategy; in this strategy, an arbitrary test of information in an initialization set is selected by substitution - meaning that the individual's information sources can be selected more than once. After a few data sets are produced, these models are trained at that time autonomously primed and depending on the type of task – i.e., regression or classification – the average or most of these predictions deliver a more accurate estimate (source: https://www.ibm.com/cloud/learn/random-forest)

Random forest algorithm

Since Random Forest uses both bagging and feature randomness to produce uncorrelated forest of decision trees, the random forest technique is an extension of the bagging method. Feature randomization provides a random subset of features, ensuring low correlation among decision trees. A significant distinction between decision trees and random forests is this. Random forests only consider a subset of the available feature splits, whereas decision trees consider all of them.

3.3.3.3 How it works

The three main hyperparameters of random forest algorithms need to be defined before training. The size of the nodes, the number of trees and the number of sample features are factors to be considered. The random forest classifier can then be used to address problems involving regression and classification. The random forest algorithm consists of a collection of decision trees and each tree in the set consists of a bootstrap sample, which is the sample of data taken from a training set with replacement. One third of the training sample is set aside as test data, which is referred to as the out-of-bag sample.

3.3.4 Random Survival Forest

RSF does not assume that the hazard ratio it is time invariant, unlike Cox PH and DeepSurv. Compared to Cox and DeepSurv, RSF offers more freedom when modeling data. In particular, the random survival forest is an ensemble tree method for the analysis of survival data with right censoring (The Annals of Applied Statistics, 2008). As is well known, developing ensembles of key learners, such as trees, can significantly improve the performance of predictions. More recently it has appeared from Breiman (2001) that component learning can be promoted by introducing randomization in the preparation of basic learning, an approach

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

called random forests. The random forest survival methodology extends Breiman's random forest strategy. In RF, randomization is presented in two forms. First a random bootstrap test of information is used to grow a tree. Second at each node of the tree, an arbitrarily selected subset of variables (covariates) is chosen as candidate factors for partition. Averaging over the trees, combined with the randomization used to grow a tree, allows RF to assume rich classes of abilities while maintaining low generalization error (The Annals of Applied statistics, 2008).

Ishwaran et al. 2008 proposed RSF, which is computed by bootstrapping data B data points from the dataset and growing B number of trees. Each tree is constructed using 70% of the provided data, while the remaining 30% is left out of the bag (Ishwaran et al. 2008). Each parent node in a tree is divided into child nodes based on the covariance that generates the largest difference in survival between nodes, given x randomly selected covariates (Iswaran et al.2008). Each tree returns a cumulative hazard function $\Lambda_b(t|_x)$ after fulfilling the required criteria. By getting the sum of every cumulative hazard function undertaken by each of the trees and dividing it by the total number of trees B, RSF returns the cumulative hazard function of the forest as follows:

$$\Lambda e(t|x_i) = \frac{1}{B}\sum_{b=1}^{B} \lambda_b(t|x_i)$$

As the exponential of the cumulative hazard function with negative sign is equal to the survival function, provided a set of variables x the RSF estimates the survival function $\hat{s}(t|x)$ as denoted below:

$$\hat{S}_{(t|x)} = \exp\left(-\hat{\lambda}_e(t|x)\right).$$

3.3.5 DeepHit

DeepHit is a multi-task network (Collobert and Weston 2008) which consists of a shared sub-network and K cause specific sub-networks (C.Lee et al., 2018). First, a single softmax layer is utilized as the output layer of DeepHit in order to ensure that the network learns the joint distribution of K competing events not the marginal distribution of each event (C.Lee et al., 2018). Second, a residual connection is maintained from the input covariates into the input of each cause - specific sub - network (C.Lee et al., 2018).

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

The shared sub – network and the k-th cause – specific subnetwork for k=1, …, K are comprised of L_s and $L_{C,k}$ fully connected layers, respectively. The shared sub – network takes as inputs the covariates x and produces as output a vector $f_s(x)$ that captures the (latent) representation that is common to K competing events (C.Lee et al., 2018).

Each cause – specific sub – network takes as inputs the pairs z = ($f_s(x)$, x) and produces as output a vector $f_{ck}(z)$, which corresponds to the probability of the first hitting time of a specific cause k (C.Lee et al., 2018). In particular, the inputs to the subnetworks include both the output of the shared network and the original covariates; this gives the sub – networks access to the learned common representation $f_s(x)$while still allowing them to learn non common part of the representation as well (C.Lee et al., 2018). If only the learned common representation were used as an input to the sub-networks, the non – common part of the representation would be lost ((C.Lee et al., 2018). The total of these outputs is a joint probability distribution on the first hitting time and event so the cause – specific sub - networks are learning the distribution for the first hitting time for each cause in parallel. The output of the softmax layer is the probability distribution y = [$y_{1,1}$,…, $y_{1,Tmax}$, …, $y_{K,1}$…$y_{K,Tmax}$]: given an employee with covariates x, an output element $y_{K,s}$ is the estimated probability $\hat{P}(s,k|x)$ that the employee will experience the event k (voluntary churn in our research) at time s (C.Lee et al., 2018). This architecture drives the network to learn potentially non – linear, even non – proportional, relation - ships between covariates and risks (C.Lee et al., 2018).

The (cause – specific) cumulative incidence function (CIF) expresses the probability that a particular event k* ∈ K occurs on or before time t* conditional on covariates x* ; as in the Fine – Gray model (Fine and Gray 1999), understanding the CIF is key to the analysis of survival under competing risks ((C.Lee et al., 2018). By definition, the CIF for the event k* is (C.Lee et al., 2018) :

$$F_{k^*}(t^*|x^*) = \text{P}\ (\text{s} \leq t^*,\ k = k^*|x = x^*) = \sum_{s^*=0}^{t^*} P\ (s = s^*, k = k^*, |x = x^*)$$

It worth mentioning here that in an unrelated way to both Cox PH and DeepSurv, DeepHit does not expect the hazard rate to be time invariant is (C.Lee et al., 2018). Furthermore, unlike the RSF presented by Ishwaran et al. (2008) , DeepHit also offers the possibility to be associated with assignments in which individuals have the probability to face not only a single event, but also non – independent events (C.Lee et al., 2018).

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

3.4 Metrics for the accuracy of the model

For the evaluation of the algorithms described above, the metrics analyzed in this subsection can be used. In particular through this thesis, we will focus on the methods used on evaluating the Random Forest Classifier algorithm used to predict employee attrition rate with IBM's dataset.

3.4.1 The score method

The score method provides information about the random forest's mean accuracy on the given data. We evaluate its performance on the training data first, and subsequently on the testing data.

3.4.2 The confusion matrix

Another tool that can be utilized to evaluate the performance of the model is by a confusion matrix. A confusion matrix shows the combination of the actual and predicted classes. Each row of the matrix represents the instances in a predicted class. It is good to measure of whether models can account for the overlap in class properties and understand which classes are most easily confused.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

4

Experimental Setup

4.1 Data

In order to perform the prediction of voluntary employee attrition with Machine learning techniques and Python and be able to investigate further its prevalent causes we made use of the data set IBM HR Analytics Employee Attrition & Performance available in Kaggle, where the last was published in 2017.

The IBM's dataset consists of a csv file with a variety of information, containing 1470 rows and 35 columns. In particular the 35 labels that characterize the workforce within the dataset provide data related to the employee's demographics (such as DistanceFromHome, Education, Marital Status, age), the employee's job position (such as Job Role, monthly income, performance rating or job involvement) and the employee's fulfillment of his or her current job (for case relationship with his coworkers). In addition, the IBM dataset demonstrates whether an employee's departure from a company occurred (SteadyLoss) and how many times a long time has passed since the employee joined the organization (AlongtimeAtCompany) and therefore is suitable for time-to event analysis.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

By reviewing further, the IBM dataset we found that 237 employees out of the 1233 in total left the company. Given that 1233 of the personnel did not experience an attrition, in terms of survival analysis right censoring of the data was applicable for 1233 employees.

Through IBM's dataset the "Attrition" feature represents each employee's decision either to leave (Yes) or to stay at the company (No) as denoted in the table below.

Figure 1. IBM dataset features

Age	Attrition	BusinessTravel	DailyRate	Department
41	Yes	Travel_Rarely	1102	Sales
49	No	Travel_Frequently	279	Research & Development
37	Yes	Travel_Rarely	1373	Research & Development
33	No	Travel_Frequently	1392	Research & Development
27	No	Travel_Rarely	591	Research & Development
32	No	Travel_Frequently	1005	Research & Development
59	No	Travel_Rarely	1324	Research & Development
30	No	Travel_Rarely	1358	Research & Development
38	No	Travel_Frequently	216	Research & Development
36	No	Travel_Rarely	1299	Research & Development
35	No	Travel_Rarely	809	Research & Development
29	No	Travel_Rarely	153	Research & Development
31	No	Travel_Rarely	670	Research & Development
34	No	Travel_Rarely	1346	Research & Development
28	Yes	Travel_Rarely	103	Research & Development
29	No	Travel_Rarely	1389	Research & Development
32	No	Travel_Rarely	334	Research & Development

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

4.2 Cleaning Process

In order to successfully complete the data cleansing process we first used the syntax df.dtypes to get the data types of each column (**Figure 2**.).

Figure 2. df.dtypes in Cleaning Process

```
#Get the column data types
df.dtypes
```

```
Age                         int64
Attrition                  object
BusinessTravel             object
DailyRate                   int64
Department                 object
DistanceFromHome            int64
Education                   int64
EducationField             object
EmployeeCount               int64
EmployeeNumber              int64
EnvironmentSatisfaction     int64
Gender                     object
HourlyRate                  int64
JobInvolvement              int64
JobLevel                    int64
JobRole                    object
JobSatisfaction             int64
MaritalStatus              object
MonthlyIncome               int64
MonthlyRate                 int64
NumCompaniesWorked          int64
Over18                     object
OverTime                   object
PercentSalaryHike           int64
PerformanceRating           int64
RelationshipSatisfaction    int64
StandardHours               int64
StockOptionLevel            int64
TotalWorkingYears           int64
TrainingTimesLastYear       int64
WorkLifeBalance             int64
YearsAtCompany              int64
YearsInCurrentRole          int64
YearsSinceLastPromotion     int64
YearsWithCurrManager        int64
dtype: object
```

Following this, we wanted to review if there were any missing or duplicate values in IBM's dataset, so we used df.isna().sum() to identify the existed empty values if any and df.isnull().values.any() as shown below (Figures 3 & 4):

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Figure 3. df.isna().sum() in Cleaning Process

```
#Count the empty (NaN, NAN, na) values in each column
df.isna().sum()
```

```
Age                         0
Attrition                   0
BusinessTravel              0
DailyRate                   0
Department                  0
DistanceFromHome            0
Education                   0
EducationField              0
EmployeeCount               0
EmployeeNumber              0
EnvironmentSatisfaction     0
Gender                      0
HourlyRate                  0
JobInvolvement              0
JobLevel                    0
JobRole                     0
JobSatisfaction             0
MaritalStatus               0
MonthlyIncome               0
MonthlyRate                 0
NumCompaniesWorked          0
Over18                      0
OverTime                    0
PercentSalaryHike           0
PerformanceRating           0
RelationshipSatisfaction    0
StandardHours               0
StockOptionLevel            0
TotalWorkingYears           0
TrainingTimesLastYear       0
WorkLifeBalance             0
YearsAtCompany              0
YearsInCurrentRole          0
YearsSinceLastPromotion     0
YearsWithCurrManager        0
dtype: int64
```

Figure 4. df.isnull().values.any() in Cleaning Process

```
[ ] #Another check for any null / missing values
    df.isnull().values.any()

    False
```

Having processed with the above steps we came to the conclusion that there were no missing values in the observations of the IBM dataset. However, the Over18, Standard Hours and EmployeeCount factors in the data all showed the same value per employee. In fact, everyone in the organization was over eighteen years old. Each employee worked forty hours per week and was counted as one person. We also identified that each employee had his/her own associate id. EmployeeNumber (id) a covariate that could be used to get a useful number. On the contrast,

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

since Over18, EmployeeCount, and Standard hours features, revealed nothing regarding the departure event, these factors were eliminated as is depicted in the figure below.

Figure 5. Removal of unneeded features

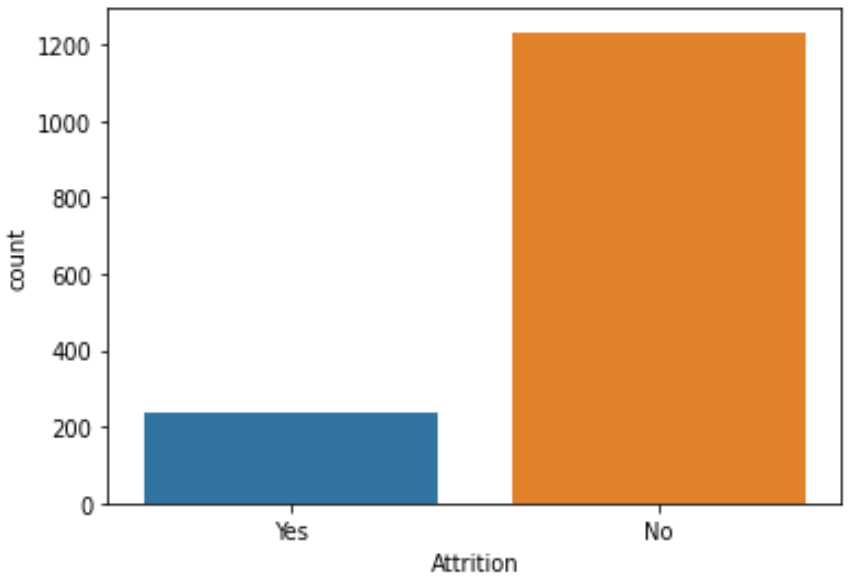
```
[ ]  #Remove unneeded columns

     #Remove the column EmployeeNumber
     df = df.drop('EmployeeNumber', axis = 1) # A number assignment
     #Remove the column StandardHours
     df = df.drop('StandardHours', axis = 1) #Contains only value 80
     #Remove the column EmployeeCount
     df = df.drop('EmployeeCount', axis = 1) #Contains only the value 1
     #Remove the column EmployeeCount
     df = df.drop('Over18', axis = 1) #Contains only the value 'Yes'
```

4.3 Dataset Visualization

After the completion of cleaning process we visualized the count of employee attrition using the syntax sns.countplot(df['Attrition']). The respective plot is depicted as follows:

Figure 6.



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Moreover we wanted to see the correlation of each column using the syntax df.corr() as depicted in the figure below.

Figure 7. df.corr()

	Age	DailyRate	DistanceFromHome	Education	EnvironmentSatisfaction	HourlyRate
Age	1.000000	0.010661	-0.001686	0.208034	0.010146	0.024287
DailyRate	0.010661	1.000000	-0.004985	-0.016806	0.018355	0.023381
DistanceFromHome	-0.001686	-0.004985	1.000000	0.021042	-0.016075	0.031131
Education	0.208034	-0.016806	0.021042	1.000000	-0.027128	0.016775
EnvironmentSatisfaction	0.010146	0.018355	-0.016075	-0.027128	1.000000	-0.049857
HourlyRate	0.024287	0.023381	0.031131	0.016775	-0.049857	1.000000
JobInvolvement	0.029820	0.046135	0.008783	0.042438	-0.008278	0.042861
JobLevel	0.509604	0.002966	0.005303	0.101589	0.001212	-0.027853
JobSatisfaction	-0.004892	0.030571	-0.003669	-0.011296	-0.006784	-0.071335
MonthlyIncome	0.497855	0.007707	-0.017014	0.094961	-0.006259	-0.015794
MonthlyRate	0.028051	-0.032182	0.027473	-0.026084	0.037600	-0.015297
NumCompaniesWorked	0.299635	0.038153	-0.029251	0.126317	0.012594	0.022157
PercentSalaryHike	0.003634	0.022704	0.040235	-0.011111	-0.031701	-0.009062
PerformanceRating	0.001904	0.000473	0.027110	-0.024539	-0.029548	-0.002172
RelationshipSatisfaction	0.053535	0.007846	0.006557	-0.009118	0.007665	0.001330
StockOptionLevel	0.037510	0.042143	0.044872	0.018422	0.003432	0.050263
TotalWorkingYears	0.680381	0.014515	0.004628	0.148280	-0.002693	-0.002334
TrainingTimesLastYear	-0.019621	0.002453	-0.036942	-0.025100	-0.019359	-0.008548
WorkLifeBalance	-0.021490	-0.037848	-0.026556	0.009819	0.027627	-0.004607
YearsAtCompany	0.311309	-0.034055	0.009508	0.069114	0.001458	-0.019582
YearsInCurrentRole	0.212901	0.009932	0.018845	0.060236	0.018007	-0.024106
YearsSinceLastPromotion	0.216513	-0.033229	0.010029	0.054254	0.016194	-0.026716
YearsWithCurrManager	0.202089	-0.026363	0.014406	0.069065	-0.004999	-0.020123

Though in terms of visualization the above result could not help us have an overview of the whole picture, so we proceeded by visualizing the correlation as follows using the below syntax:

Figure 8.

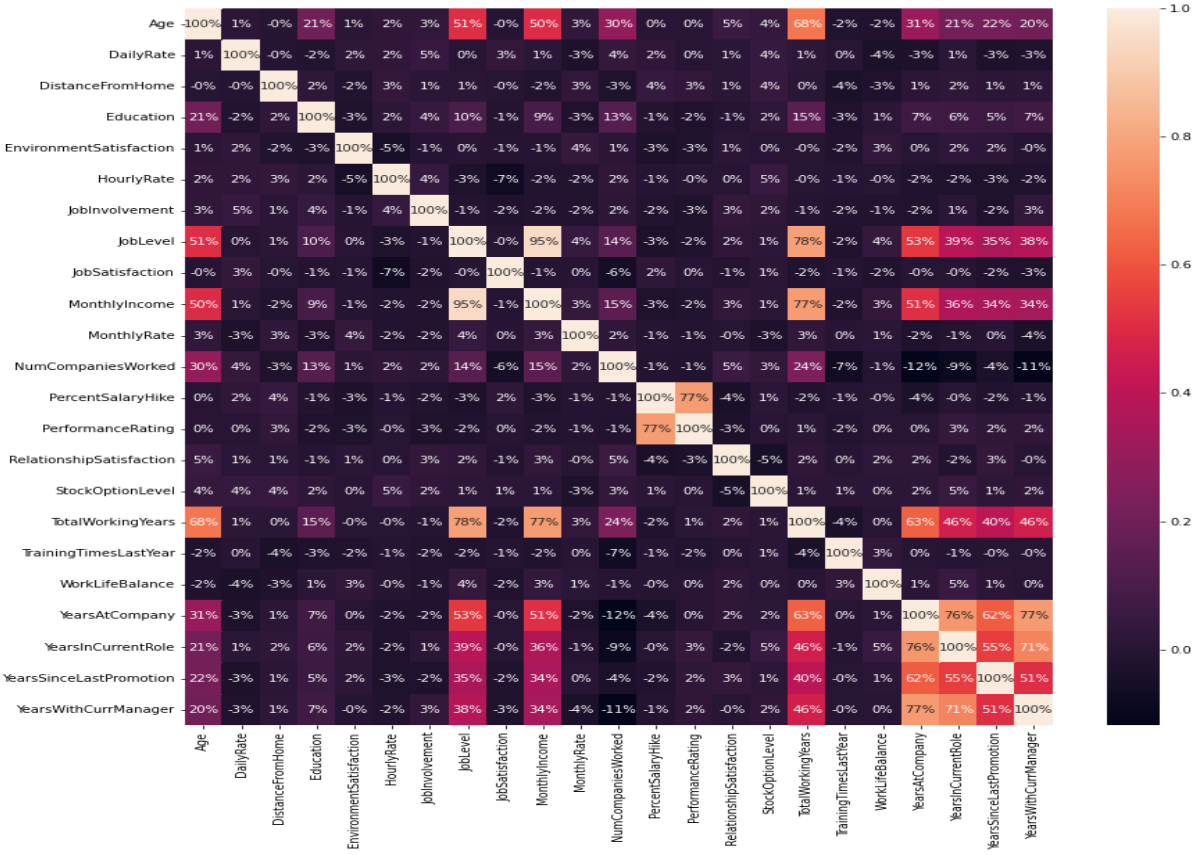
```
#Visualize the correlation
plt.figure(figsize=(14,14))  #14in by 14in
sns.heatmap(df.corr(), annot=True, fmt='.0%')
```

We observed that age column had a positive correlation of sixty percent with the total working years of an employee within the organization which was a logical outcome since the longer someone is employed within a company the older, he is getting. Also, job level had 78% correlation with total working years which indicates that the longer someone is working to a specific position the better value of job he/she provides. Finally, worth mentioning that monthly income had 77% correlation with total working years indicating that the longer someone is employed within the organization is associated to a proportional increase of his salary as well

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

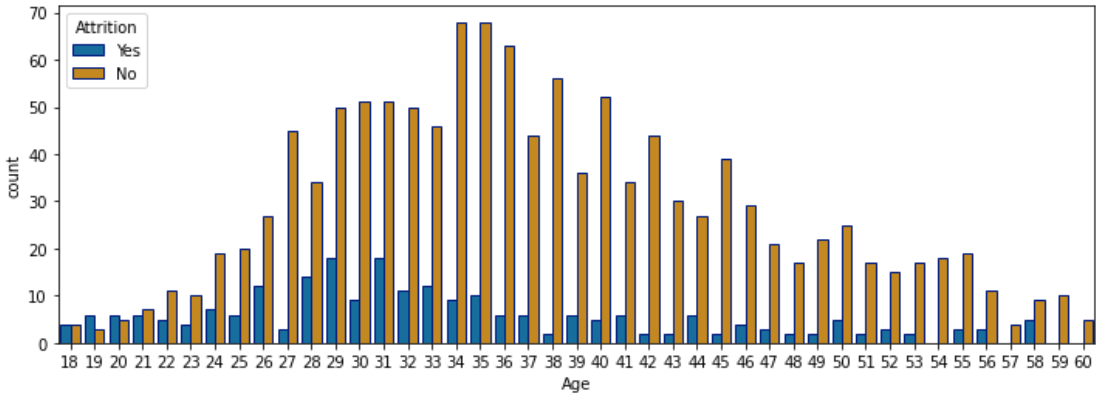
as 95% with the job levels since in accordance with the seniority of each role the salary is higher respectively. All the above observations are depicted in Figure 8 below.

Figure 9: Correlation



As a next step we intended to illustrate the employees who left the company based on age. What we noticed was that the employees with the highest attrition rate were those between 29 & 31 while the age with the best retention was 34 & 35.

Figure 10.



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

4.4 Model Building

In this section we do present how we prepared the data for the model. First, we proceeded with transforming non-numeric columns to numerical columns. The code used for this purpose is depicted below:

Figure 11.

```
#Transform non-numeric columns into numerical columns
from sklearn.preprocessing import LabelEncoder

for column in df.columns:
        if df[column].dtype == np.number:
            continue
        df[column] = LabelEncoder().fit_transform(df[column])
```

Afterwards we created a new column as a storage for the age's values. This was only in order to place the age values at the end of the dataset. At that point, we would remove the respective column containing age from the front of the dataset so that the target feature to be first. Finally, we would show the new data set. The code and the amended order of the dataset reflecting the steps described is denoted below:

Figure 12.

```
#Create a new column at the end of the dataframe that contains th
e same value
df['Age_Years'] = df['Age']
#Remove the first column called age
df = df.drop('Age', axis = 1)
#Show the dataframe
Df
```

	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EnvironmentSatisfacti
0	1	2	624	2	0	1	1	
1	0	1	113	1	7	0	1	
2	1	2	805	1	1	1	4	
3	0	1	820	1	2	3	1	
4	0	2	312	1	1	0	3	
...	...	...	...	...	...	...	...	
1465	0	1	494	1	22	1	3	
1466	0	2	327	1	5	0	3	
1467	0	2	39	1	3	2	1	
1468	0	1	579	2	1	2	3	
1469	0	2	336	1	7	2	3	

1470 rows × 31 columns

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Following the above we had to split the dataset into independent 'X' and dependent 'Y' variables the respective code is denoted as follows:

Figure 13.

```
#Split the data into independent 'X' and dependent 'Y' variables
X = df.iloc[:, 1:df.shape[1]].values
Y = df.iloc[:, 0].values
```

Then, we proceeded with splitting the data set into 75% training and 25% testing as depicted below:

Figure 13.

```
# Split the dataset into 75% Training set and 25% Testing set
from sklearn.model_selection import train_test_split
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_si
ze = 0.25, random_state = 0)
```

Below we present how we utilized the Random Forest Classifier to learn from the training data and check the accuracy of the model:

Figure 14.

```
from sklearn.ensemble import RandomForestClassifier
forest = RandomForestClassifier(n_estimators = 10,
criterion =   'entropy', random_state = 0)
forest.fit(X_train, Y_train)
```

As our next step we wanted to measure the exact accuracy of our model, we did this with the use of the following syntax: *forest.score(X_train, Y_train)*

The model was characterized by 97.9 % accuracy on the training data, this is denoted via the following figure:

Figure 15.

```
[ ]  forest.score(X_train, Y_train)

     0.9791288566243194
```

In addition, we showed the confusion matrix and accuracy for the model on the test data. Just to highlight at this point that the classification accuracy is defined as the ratio of the correct predictions to the total predictions made. The code for this step is depicted through the figure below:

Figure 15.

```
#Show the confusion matrix and accuracy for  the model on the tes
t data
#Classification accuracy is the ratio of correct predictions to t
otal predictions made.
```

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

```
from sklearn.metrics import confusion_matrix

cm = confusion_matrix(Y_test, forest.predict(X_test))

TN = cm[0][0]
TP = cm[1][1]
FN = cm[1][0]
FP = cm[0][1]

print(cm)
print('Model Testing Accuracy = "{}!"'.format(  (TP + TN) / (TP +
 TN + FN + FP)))
print()# Print a new line
[[309   1]
 [ 49   9]]
Model Testing Accuracy = "0.8641304347826086!"
```

In accordance with the result of the above, we identified that the model identified with 86.41% accuracy the employees that left the company.

Afterwards, we wanted to review which were considered the most important features for the model and had a major role consequently to the attrition rate of the employees within the organization. In order to denote this, we used the code depicted below:

Figure 16.

```
#Return the feature importances (the higher, the more important t
he feature).
importances = pd.DataFrame({'feature':df.iloc[:, 1:df.shape[1]].c
olumns,'importance':np.round(forest.feature_importances_,3)}) #No
te: The target column is at position 0
importances = importances.sort_values('importance',ascending=Fals
e).set_index('feature')
importances
```

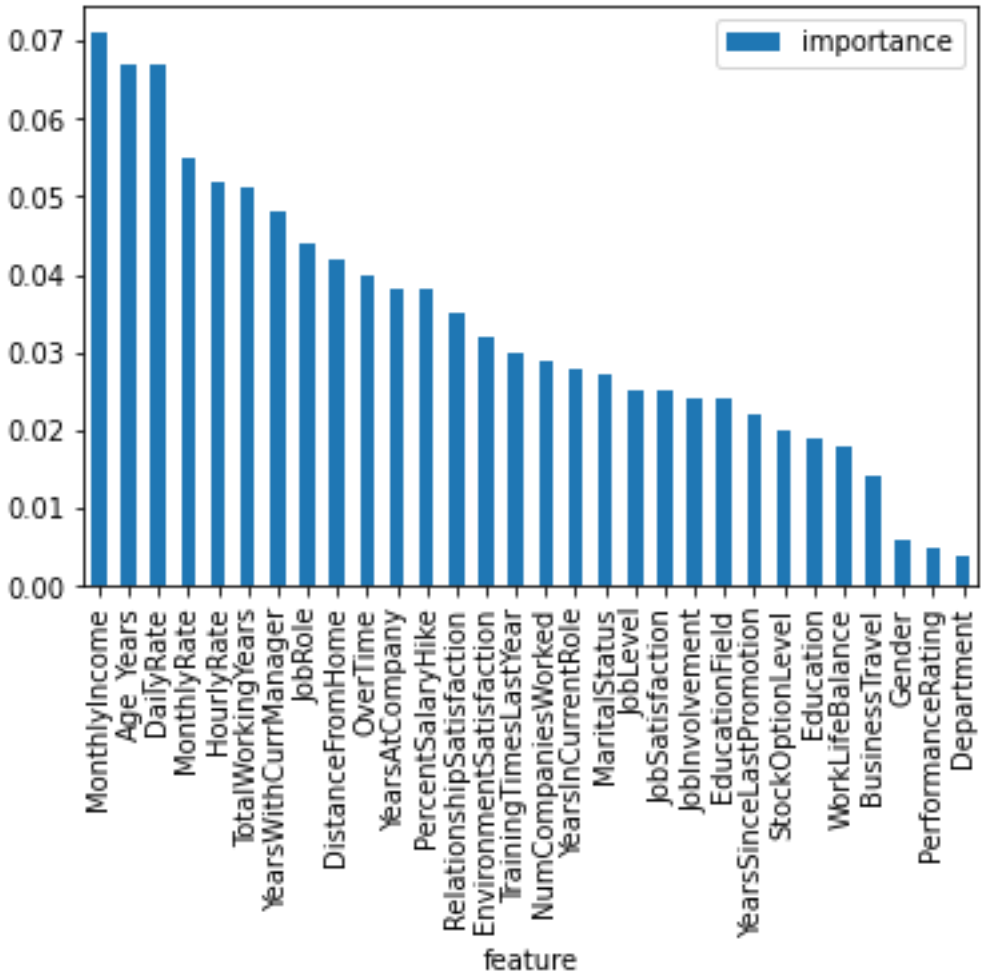
feature	importance
MonthlyIncome	0.071
Age_Years	0.067
DailyRate	0.067
MonthlyRate	0.055
HourlyRate	0.052
TotalWorkingYears	0.051
YearsWithCurrManager	0.048
JobRole	0.044
DistanceFromHome	0.042
OverTime	0.040
YearsAtCompany	0.038
PercentSalaryHike	0.038
RelationshipSatisfaction	0.035
EnvironmentSatisfaction	0.032
TrainingTimesLastYear	0.030
NumCompaniesWorked	0.029
YearsInCurrentRole	0.028
MaritalStatus	0.027
JobLevel	0.025

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Since there were a lot features to look at, we wanted to see a visualization of the data. In order to visualize this, we used the code as denoted below:

Figure 17.

```
[ ]  #Visualize the importance
     importances.plot.bar()
```



As we can observe from Figure 17. monthly income seems to be the most important feature followed by the age of the individual, the daily rate, and the monthly rate. It is indisputable the fact that income plays a vital role in the attrition rate of an organization since someone's salary is very important and can play a key role in their decision to leave their current job.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

5

Evaluation

In this part we are going to present two evaluation experiments that have been conducted regarding the employee attrition analysis. The first one describes the contribution of the CoxRF algorithm since the last has the advantage by combining the statistical results of survival analysis with the help of assembly learning to reduce the problem called conservative supervised binary classification for an event centered perspective. The second one refers to an improved machine learning-based employees attrition prediction framework with emphasis on feature selection. In particular, the second experiment being analyzed and evaluated, presents a three – stage framework for the prediction of the turnover rate of employees.

5.1 Talent Flow Employee Analysis based Turnover Prediction on Survival Analysis (Sumathi K. et al.,2021)

In this experiment to help structure survival information from censored information, the terms "event-person" and "time event" were coined. The CoxRF was interrelated to a number of baseline approaches using an original dataset of China's largest technological network. The outcome showed that it is a good attrition interpreter. The following are some of the findings that have been made: i) employee turnover varies by industry, with the IT sector having a slightly higher rate than the government sector; ii) gender plays a major role if it is a woman after marriage, some are relieved from work and other factors as well; iii) a person with good academic records can work more efficiently than another with low; iv) GDP plays an important role in company and employee turnover in the current situation, which has been overlooked in previous studies; v) and the final point is that the wage increase they are implementing is one of the reasons we are losing a terrific employee.

This experiment used survival analyzing and algorithms of machine learning to predict the person will quit his ongoing work at a time (t) given his attrition events on past, on working job information, information on social networks and a specific time t. The CoxRF was

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

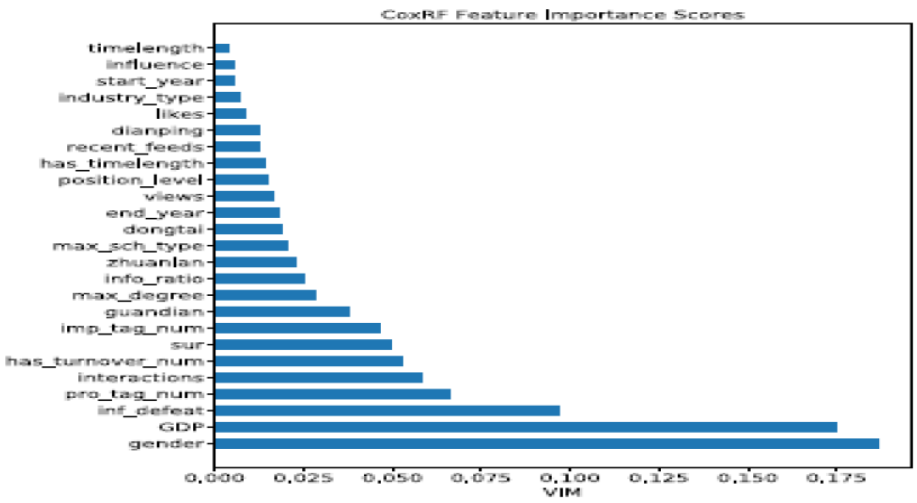
compared with baseline algorithms. Eventually, Kaplan – Meier analysis was used to identify factors influence employee turnover behavior. Initially employee A's were separated in two events A1 and A2 categories. Then, rather than of focusing on the calendar year, the absolute time was converted to relative time or in other words the numbers 0-6 were used to represent the years 2000 to 2006 with the length of work time as the key factor. The A1 and A2 files were processed as "case person" which indicated that the event was broken into sections based on who was involved. A year event on the other hand, was a relative time period spanning from 0 to 6 that was part of the "time event" definition, which divided time by event.

Following the data cleaning there were 287,229 job records. Based on the start and the end times of an experience work, the label was set to consider the user to finally abandon the task. If the user filled in the start time and left the end time blank, the label was set to 0 and the user was in a work state. The data set was randomly divided into training and testing sets in a 7:3 ratio. The accuracy recall F1 measure was used and AUC metrics to evaluate the model.

5.1.1 Results

Gender: In figure 18, the feature important scores are displayed, followed by the determined average value of the gini index score, which is then normalized. Gender and lucrative indication are two variables that can raise their scores to 0.15, with gender being a category variable. The Kaplan – Meier method was used to determine the survival rate and group the plot survival curves based on the differences between groups. The survival curves of both genders were similar in shape.

Figure 18. Sumathi K. et al.,2021



However, when the details of the curve were examined, it was found that female survival rates were consistently lower than male survival rates, meaning that after working for the same

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

number of years, more females were likely to quit. The female endurance curve dropped from 200 to 400 (month to event) for the female category.

Industry: The dataset used also included detailed information about the sectors in which employee works, the last were 18 categories in total. It was shown that job changers were mostly IT employees while government ones were more likely to stay in government sectors only. Other companies' turnover rates could reach 20% (80% below survival probability) within 20 months of starting work, an all have similar drop patterns and survival curves. The government sector took three years to reach this level.

Educational Background: in order to explore how educational background effects employee turnover, three kinds of employees were divided based on their highest academic degree at a university level, namely project 985 university, project 211 university, and other.

To summarize in this experiment the Cox RF approach was proposed for predicting employee turnover on the survival analysis. Random forest bagging ensemble learning, and survival analysis were combined in CoxRF. The higher the number of events and passages, the higher the survival rate. Meanwhile this work was renamed as standard supervised binary classifier, and it was compared to all other algorithms. Moreover, time event and event – person concepts for building data survival and maxi missing using censored data were suggested. It was discovered that sex (gender) had a substantial impact on business turnover, with female candidates having a higher turnover rate than male candidates for the same amount of time worked. Furthermore, it was shown that the turnover of different industries varies.

5.2 An improved machine learning – based employees attrition prediction framework with Emphasis on Feature selection

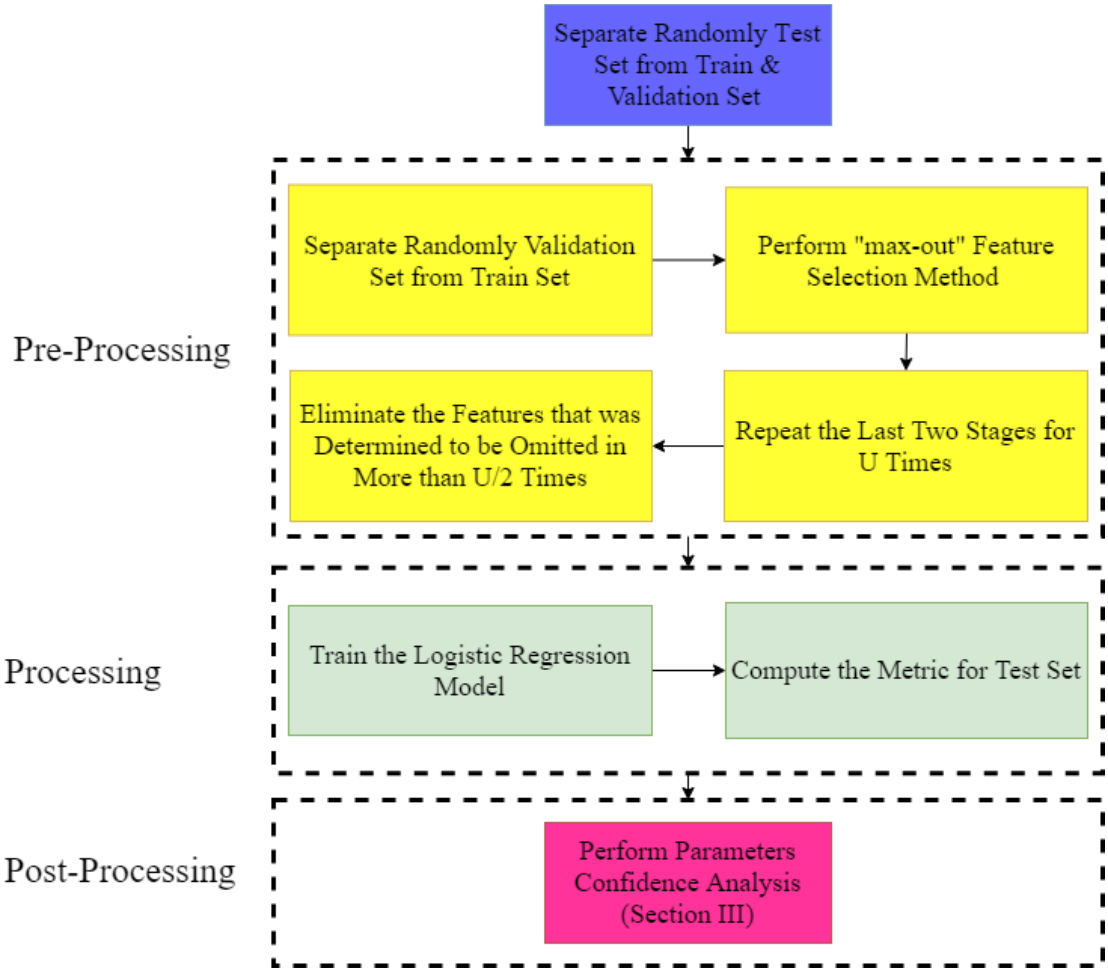
This experiment presents a three – stage (pre-processing, processing, post – processing) framework for attrition prediction. Again, the IBM data set was used to implement this approach. The importance of each feature in the logistic regression model was represented by its coefficient in the prediction of attrition. The F1-score performance measure has improved as a result of the findings due to the feature selection process of "maxing out". Finally, the parameters' validity was confirmed. Multiple bootstrap datasets were used to train the model. The average and standard deviation of the parameters examined were then calculated to see if they had a high level of confidence and were stable. The model's small standard deviation indicated that it was stable and more likely to succeed.

This experiment proposed an attrition prediction task that addressed all three stages of preprocessing, process, and post pre-processing. The pre-processing stage begun with the

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

"max-out" technique which is a novel feature selection strategy for improving the performance of the attrition prediction classifier. Then, for the processing stage, a logistic regression model was trained for the new collection of features. Following that, in the post-processing stage, confidence analysis was used to measure how certain we were about our model's parameters. Finally, IBM attrition data was used to validate the methodology. Figure 19 depicts the proposed framework's overall structure. Pre-processing, processing, and post-processing stages are represented by yellow, green, and red blocks in this diagram. The major goal of these phases is to ensure that the model can generalize correctly.

Figure 19. Najafi-Zangeneh et al., 2021



The "max - out "algorithm feature selection, which belongs to the wrapping category, was designed based on the nature of this problem's feature set, which comprises both binary and continuous features. Algorithm 1 encapsulated the algorithm. First all subsets of n-m features were trained using this technique. The feature set was picked from the subgroup with the most

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

significant measure. For each new collection of features, the process was repeated. The feature set was not altered any more when the metric became less than the preceding stage. The approach was substantially faster than examining all possible combinations of features because the model was only trained for a subset of all available features. When m is 1, the algorithm is known as 1-max-out and when m is 2 the algorithm is known as 2-max-out. Backward feature selection is the 1-max-out algorithm. However, in some circumstances, combining m characteristics may improve performance. Nonetheless, each of them may or may not play a substantial effect in classification performance. As a result, 1-max-out may incorrectly destroy these functionalities one by one. In these situations, m-max-out (m>1) outperforms 1-max-out. Given f as the number of initial features, the m-max-out is on the order of O(fm). As a result, selecting an appropriate m is also influenced by the computing resources available.

As previously mentioned, this experiment was again based on the IBM attrition data set. For each employee there were 35 columns in this dataset. The classifier's target output was attrition which was one of these columns. The remaining 34 columns are called features. "Standard hours" and "employee count" were two of these attributes that were shared by all employees. As a result, in this experiment like the one previously runed in this thesis using google collabs, in the Experimental setup section, these two attributes were not included in the features. Other characteristics included "age", "education field", "department", "daily rate", "job involvement", "job level", "monthly income", "monthly rate", "performance", "job role", "job satisfaction", "marital status", "percent salary hike", and "years with the current manager". In this experiment the data set was initially partitioned into the training and validation set and test set in order to determine which attributes were the most essential. The validation set was then separated from the series of exercises. In order to determine, the 1-max-out method was used here. Which features should be left out? Following then, the procedure of randomly dividing the participants begun. The validation set and 1-max-out procedure were repeated. The features were finalized after seven revisions. That were found to have been omitted more than four times were removed.

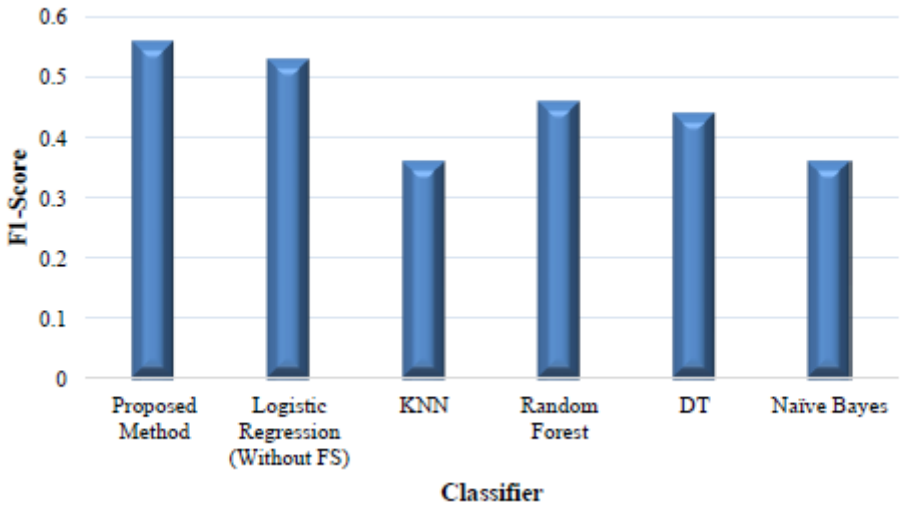
This technique yielded that "Education Field_HR", "Monthly income","Gender female" , "hourly rate", "Department research and development", "Over18_yes", "Education", "Job level", "Department_Research & Development", "performance rating", "Job Role_Manufacturing Director", "Monthly rate", "Education Field other", "Years at company", "Departments_Sales", "Over Time_No", "Education Field_Marketing" were left out. These characteristics weren't necessarily the least important for predicting attrition. Because they were totally associated with other aspects in the dataset, some of them were chosen to be eliminated. "Gender female" for example was one without the "Gender male" attribute. Being eliminated for categorical features that were translated to binary features means that being in

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

this category has no effect on the probability of attrition. Figure 20. shows the value of the coefficients for each feature. “Years since previous promotion”, “overtime” and working as a sales representative, according to the coefficients, were the most influential factors for an employee to resign from their job. As any of these variables rise, the likelihood that the employee would resign rises. As a result, the job’s value rises. Working as a research director, on the other hand, entailed “total working time”. The most influential factors were “years with present manager” and “job involvement” variables that influence an employee’s decision to stay with the organization.

Figure 20. Najafi-Zangeneh et al., 2021

Feature	Coef.	Feature	Coef.	Feature	Coef.
Age	−0.776	Environment Satisfaction	−1.174	Education Field_Life Sciences	−0.181
Daily Rate	−0.738	Business Travel_Travel_Frequently	0.810	Education Field_Technical Degree	0.341
Distance From Home	1.004	Percent Salary Hike	−0.642	Training Times Last Year	−0.835
Job Involvement	−1.536	Number Companies Worked	1.375	Job Role_Laboratory Technician	1.009
Relationship Satisfaction	−0.701	Job Satisfaction	−1.116	Job Role_Sales Executive	0.762
Stock Option Level	−0.255	Total Working Years	−1.887	Marital Status_Divorced	−0.728
Work Life Balance	−0.852	Years with Current Manager	−1.615	Job Role_Manager	−0.670
Years in Current Role	−1.287	Years Since Last Promotion	2.925	Job Role_Sales Representative	1.483
Job Role_Health care Representative	−0.333	Gender_Male	0.606	OverTime_Yes	1.996
Job Role_Human Resources	0.463	Job Role_Research Scientist	−0.096	Job Role_Research Director	−2.178
Marital Status_Single	0.755	Constant	2.176		



The approach used in order to check the confidence value for each coefficient from the original dataset 300 bootstrap datasets were generated. The model was then trained for each dataset separately. The following table shows the average, standard deviation, and coefficient of variations for all coefficients. The standard deviations indicated the level of confidence on

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

average. We could be more sure in fields where the coefficient of variations was low. For instance, the most faith coefficient was “overtime” feature.

Figure 21. Najafi-Zangeneh et al., 2021

Feature	Ave.	Feature	Ave.	Feature	Ave.
	Std.		Std.		Std.
	CV		CV		CV
Age	−0.786	Daily Rate	−0.753	Distance From Home	1.007
	0.320		0.221		0.214
	0.407		0.293		0.212
Environment Satisfaction	−1.181	Business Travel_Travel_Frequently	0.824	Job Involvement	−1.572
	0.162		0.147		0.235
	0.137		0.178		0.149
Job Satisfaction	−1.131	Number Companies Worked	1.390	Percent Salary Hike	−0.650
	0.166		0.219		0.228
	0.146		0.157		0.350
Education Field_Life Sciences	−0.182	Relationship Satisfaction	−0.717	Stock Option Level	−0.262
	0.122		0.172		0.264
	0.670		0.239		1.007
Total Working Years	−1.962	Training Times Last Year	−0.850	Education Field_Technical Degree	0.358
	0.503		0.255		0.213
	0.256		0.3		0.595
Work Life Balance	−0.869	Years in Current Role	−1.308	Years Since Last Promotion	2.952
	0.220		0.403		0.361
	0.253		0.308		0.122
Years with Current Manager	−1.630	Gender_Male	0.358	Job Role_Health care Representative	−0.329
	0.409		0.213		0.340
	0.250		0.595		1.03
Job Role_Research Director	−2.147	Job Role_Human Resources	0.450	Job Role_Laboratory Technician	1.021
	0.371		0.351		0.208
	0.172		0.78		0.203
Job Role_Manager	−0.622	Marital Status_Divorced	−0.755	Job Role_Research Scientist	−0.101
	0.280		0.170		0.192
	0.450		0.225		1.9
Job Role_Sales Executive	0.776	Job Role_Sales Representative	1.500	OverTime_Yes	2.029
	0.200		0.264		0.124
	0.257		0.176		0.061
Marital Status_Single	0.764	Constant	2.215		
	0.170		0.403		
	0.222		0.182		

The variation of the parameter over all bootstraps could be also graphically demonstrated using box plots of the coefficients. The variation of coefficients related with Figure 22 was depicted with the most important characteristics, which were addressed in the preceding part. The years since the last promotion’s coefficient took a value between 2 and 4 in the following plot for all the bootstrap training datasets. As a result, we could have faith in it. Attrition was a significant consequence of the “Over Time-Yes” feature’s coefficient barely varies. Thus, we could be certain of the coefficient’s value. On the other hand, the worth of the coefficient for “Years with current manager” fluctuated a lot. Therefore, we could not be sure about this parameter. The last, however, was negative in all the bootstrap datasets. Consequently, it could be deduced that this feature had a positive impact on keeping employees within the organization.

To summarize, the goal of this experiment was to demonstrate a machine learning model for forecasting staff attrition. The initial step was to offer a feature selection strategy for lowering the dimension of the feature space. Then, for the aim of prediction, a logistic model was trained.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

When compared to logistic approaches, the results showed that the proposed feature selection improved the predictor's performance. The model revealed that the most commons reasons for quitting the work were "years after the last promotion", "Overtime – Yes", "Job Role Sales Representative", and "Number of companies worked". Larger values for these characteristics indicated a higher likelihood of attrition. In contrast, "total working years", "years with present boss" and "job participation" were the most important factors in deciding whether or not to stay with the organization. In particular 300 hundred bootstrap datasets were created to test whether the parameters were valid. A model was created for each of those. The coefficients of each attribute were then statistically analyzed. In general, the coefficient variation was acceptable. Variations in parameters related with the most influential traits were minor. Thus, we were confident that the aforementioned characteristics were the most important in forecasting attrition.

In contrast to earlier research this experiment proposed a three – stage approach for constructing a precise employee attrition model, including pre-processing and post processing as well as for determining the model's parameter's validity. The m-max-out algorithm was introduced at the pre-processing stage for feature selection. The 1-max-out (which is a specific situation in which m is equal to one) was employed in this experiment due to the limitations of compute devices. In case of greater available computational resources, a larger m could also be used. The validity of logistic regression model's parameters for attrition prediction was tested by looking at how they changed when trained over multiple bootstrap datasets. These stages of preprocessing and postprocessing could be utilized to create accurate and reliable models for any generic situation. Any set of feature sets, including binary and continuous features, could be employed with the max-out feature selection approach. Statistical study of the model's parameters over many bootstraps could refer whether we had confidence in the model for any type of parametric Machine Learning model.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

6

Conclusions & Future Research

Employee turnover is linked to increased costs. Managers could implement retention measures to keep employees from leaving. By addressing turnover costs and identifying employees who are prone to leave, these retention tactics can be implemented. Only voluntary, preventable, and dysfunctional turnover can be avoided with retention methods.

The aim of this study was to determine whether survival analysis and machine learning algorithms could accurately predict voluntary departure of employees. The findings indicated that both survival methods and machine learning algorithms as well as the combination of those two, could accurately predict intentional behavior.

Having conducted our research, the advantages as well disadvantages of using survival analysis are depicted as follows:

Advantages

1. Better utilization of macroeconomic data that change over time
2. Estimation of medium – term cancellation and non-renewal of an employee at the end of the period in turn and simultaneously
3. It is considered not just whether an event will be discontinued, but also when.
4. Using panel data, it provides a dynamic perspective and enhance the static view generated from snapshot data

Disadvantages

1. The implementation of the model is more complicated than that of a binary model
2. Macroeconomic variables that change throughout time are more difficult to forecast than retention

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Also, in order this research to be conducted we used and presented machine learning algorithms and showed their contribution in predicting voluntary turnover of employees. In particular we used Random Forest Classifier to predict the churn rate of employees in IBM's data set. The outcome proved our model was characterized by 97.9 % accuracy on the training data and with an 86.41% accuracy on the test data (attrition feature - employees that left the company). Moreover, we presented in detail two experiments have been made regarding the prediction of voluntary turnover; the first using survival analyzing and algorithms of machine learning to predict the person will quit his ongoing work at a time (t) while the second utilizing improved machine learning-based employees attrition prediction framework with emphasis on feature selection.

In that regard, the following suggestions / assumptions, which need to be empirically verified, can be made on the way forward in terms of ML methods, and these can be enriched with future research topics:

- Obtain more knowledge about the unknown future values of the data rather than the historical values, then optimize/ learn as much as possible using these values.
- Prior employing machine learning algorithms, deseasonalize the data. As a result, it will be easier to learn because the computing time required to arrive at ideal weights will be reduced.
- Use a slip simulation to gather as much as possible about future values and the uncertainty surrounding them and learn more effectively how to reduce them.
- Cluster the data into multiple homogeneous categories and/or types of data, then construct machine learning methods to extract them effectively.
- Avoid overfitting because it is not clear whether ML models can effectively distinguish noise from the data model.
- Preprocessing can be automated to eliminate the need for the user to make additional judgments.
- Allow the estimation of uncertainty in point forecasts as well as the building of confidence intervals around such forecasts.

Through our research it was denoted that ML models themselves have a lower prediction accuracy than combined with statistical methods such as survival analysis. So, we are optimistic that more attempts will be implemented resulting of their significant improvement in accuracy over time.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Bibliography

[AHF+21] Salah Al-Darraji, Dhafer G. Honi, Francesca Fallucchi, Ayad I. Abdulsada, Romeo Giuliano, Husam A. Abdulmalik. Employee Attrition Prediction Using Deep Neural Networks, 2021.

[N20] Fredrick Norman. Predicting employee attrition with machine learning on an individual level, and the effects it could have on an organization, 2020.

[JJP20] Praphula Kumar Jain, Madhur Jain, Rajendra Pamula. Explaining and predicting employees' attrition: a machine learning approach, 2020.

[KSC+18] Jared L. Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang and Yuval Kluger. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network, 2018.

[BCS03] MJ Bradburn, TG Clark, SB Love and DG Altman. Survival Analysis Part II: Multivariate data analysis – an introduction to concepts and methods., 2003.

[KP02] Kalbfleisch JD and Prentice RL. The Statistical Analysis of Failure Time Data, 2002.

[KP97] Klein and Moeschberger. Survival Analysis: Techniques for Censored and Truncated Data, 1997.

[L79] Lagakos SW. General right censoring and its impact on the analysis of survival data, 1979.

[L03] Lawless JF. Statistical Models and Methods for Lifetime Data. Wiley, New York., 2003.

[C18] Yen-Chi Chen. Lecture 5: Survival Analysis, 2018.

[MML+99] Paula C. Morrow, James C. McElroy, Kathleen S. Laczniak, James B. Using Absenteeism and Performance to Predict Employee Turnover: Early Detection through Company Records, 1999.

[YHC21] NESRINE BEN YAHIA, JIHEN HLEL, RICARDO COLOMO-PALACIOS . From Big Data to Deep Data to Support People Analytics for Employee Attrition Prediction, 2021.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

[MS21] Norsuhada Mansor, Nor Samsiah Sani. Machine Learning for Predicting Employee Attrition, 2021.

[PB+21] Madara PRATT, Mohcine BOUDHANE, Sarma CAKULA. Employee Attrition Estimation Using Random Forest Algorithm, 2021.

[DA13] Alao D. & Adeyemo A. B. ANALYZING EMPLOYEE ATTRITION USING DECISION TREE ALGORITHMS, 2013.

[GL10] Matteo Gagliolo, Catherine Legrand. Algorithm Survival Analysis, 2010.

[MO+17] Rafa Madariaga, Ramon Oller, Joan Carles Martori. Discrete choice and survival models in employee turnover analysis, 2017.

[BG21] Nilasha Bandyopadhyay, Anil Jadhav. Churn Prediction of Employees Using Machine Learning Techniques, 2021.

[NB+21] Thalapaneni Penchala Naidu, Dr. B. Sankara Babu, K Saikumar, Bhavana Godavarthi, Paparao Nalajala. Detection And Identification Of An Employee Attrition Using Machine Learning Algorithm, 2021.

[RN+20] M. Ravi, A. Nirmai, A. Krishitha, CH. Madhan Mohan Reddy. Prediction of Employee Attrition using Random Forest Classifier Technique, 2020.

[GW+18] Xiang Gao, JunhaoWen, Cheng Zhang. An Improved Random Forest Algorithm for Predicting Employee Turnover, 2020.

[P16] Christopher E. Penney. A survival analysis of ADM (Materiel) workforce attrition, 2016.

[B15] Wioletta Grzenda. Estimation of Employee Turnover with Competing Risks Models, 2015.

[MS+17] Spyros Makridakis, Evangelos Spiliotis, Vassilios Assimakopoulos. Statistical and Machine Learning forecasting methods: Concerns and ways forward, 2017.

[BS18] Kashyap Bhuva, Kriti Srivastava. Comparative Study of the Machine Learning Techniques for Predicting the Employee Attrition, 2018.

[SE21] Praveen Ranjan Srivastava, Prajwal Eachempati. Intelligent Employee Retention System for Attrition Rate Analysis and Churn Prediction: An Ensemble Machine Learning and Multi-Criteria Decision-Making Approach, 2021.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

[YI21] Shenghuan Yang, Md Tariqul Islam. IBM Employee Attrition Analysis, 2021.

[PA16] Rohit Punnoose, Pankaj Ajit. Prediction of Employee Turnover in Organizations using Machine Learning Algorithms: A case for Extreme Gradient Boosting, 2016.

[BN+21] Sumathi K., Balakrishnan D., Naveen V., Hariharan P., Rahul Iniyan M. Talent Flow Employee Analysis based Turnover Prediction on Survival Analysis, 2021.

[FW13] Luyang Fu, Hongyuan Wang. Estimate Attrition Using Survival Analysis, 2013.

[NS+21] Saeed Najafi-Zangeneh, Naser Shams-Gharneh, Ali Arjomandi-Nezhad and Sarfaraz Hashemkhani Zolfani. An Improved Machine Learning-Based Employees Attrition Prediction Framework with Emphasis on Feature Selection, 2021.

[R19] Omar RAJÂA. A Review of the Literature on Employee Turnover, 2019.

[LA+12] Choi Sang Long, Musibau Akintunde Ajagbe, Khalil Md Nor and Ebi Shahrin Suleiman. The Approaches to Increase Employees' Loyalty: A Review on Employees' Turnover Models, 2012.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

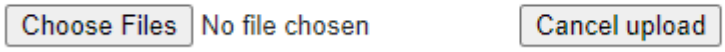
Appendix: Running the Simulation Environment

In order to run the simulation environment, you need to have access to Google Collabs so no need for a specific runtime environment in your computer. The zip file that contains IBM's data can be downloaded from Kaggle, via the link depicted below:

https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

Once you download the file, unzip it, and upload the csv. file to the respective section as denoted below:

```
#Load the data
from google.colab import files # Use to load data on Google Colab
uploaded = files.upload() # Use to load data on Google Colab
```



ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ