# EMPLOYEE RETENTION PREDICTION IN CORPORATE ORGANIZATIONS USING MACHINE LEARNING METHODS

**Khaled Alshehhi, Abu Dhabi School of Management**
**Safeya Bin Zawbaa, Abu Dhabi School of Management**
**Abdullah A Abonamah, Abu Dhabi School of Management**
**Muhammad Usman Tariq, Abu Dhabi School of Management**

## ABSTRACT

*Employee Retention is the capability of an organization to maintain its employees. The concept is emerging as a key setback to organizations. Payments, organization culture, job satisfaction, remuneration, and flexibility impact the rate of retention for any organization or company. Employee Retention is also an essential function of Human Resource Management. Unless there is a thoughtful and serious effort from the management towards this direction, the competitors within the industry are likely to snatch and attract the talent which another company had nurtured over a period of time. Appropriate approaches for the formulation and implementation of employee retention approaches are a skill and needs to be prioritized by the management. The paper focuses on providing the prospective reasons why employees leave their jobs. The paper also examines the talents which companies want to develop and maintain to predict employee training. Similarly, the paper identifies the immediate productive scrutiny methods and implementations in a practical scenario, and who is eligible to be retained in the company and have a productive career path. Lastly, the paper identifies the essential factors for developing the predictive model. To achieve these, the research employed a survey in collecting primary data. The study surveyed the employees to identify the attributes and factors that are essential in predicting retention rates and improving retention of employees. The research conducted experiments with a machine-learning algorithm on the dataset, which increased the accuracy of the research outcomes. These assisted the research in identifying the employee retention rate. Training count emerged as the top predictor of employee retention. Thus, this research suggests that the company should strive to train more employees since those who have attended more training are retained.*

**Keywords:** Employee Training, Employee Retention, Training, Machine Learning, Support Vector, Random Forest

## INTRODUCTION

Challenges linked to employee retention are emerging issues in contemporary corporate organizations. The main problem lies in the vital workforce management and is likely to pose challenges to companies in the near future. Companies are likely to adapt to the various forms of organizational behavior concerning the realities of the current work environment, whose success and longevity relies on creativity, innovativeness, and flexibility. Employee retention is a multifaceted ideology due to a lack of a specific approach that is applicable in securing employees within the organization.

Kumar & Mathimaran (2017) defined retention as the mandate to continue engaging in business activity with a specific company on an on going basis. Various current and ample

descriptions for the idea of retaining employees rely on commitment, liking, identification, readiness to recommend, trust, and repurchase intentions, with the first four being behavioural intentions. The last two are emotional-cognitive retention.

The human resource in every organization is perceived to be the sole determinant of competitive advantage and the most valuable asset in every organization. For every organization, competent employees are the backbone. Therefore, organizations need to meet modern challenges by providing training opportunities for the employees to support the objectives of the organizations (Kumar & Mathimaran, 2017). When organizations ignore development and training programs, there is a high possibility of more staff turnover compared to the organizations conducting such programs (Kumar & Mathimaran, 2017). Training of employees yields outstanding outcomes to the growth and development of organizations. However, the organization has to bear the cost when an employee leaves the organization. Such expenses are almost equal to the benefits of the employee, and more than a year's salary.

According to Ramos (2019), the higher the organizational commitment, the lesser the employee turnover. Organizational commitment is an essential variable concerning a more extended stay and employee retention. Committed employees exhibit more secondary intentions to quit. Such employees exert the maximum efforts of fulfilling the assigned tasks for the benefit of the organization. Both the employer and the organization are primarily concerned with the appropriate commitment of the employees. Employee commitment is an essential factor that is vital to improving organizational performance. Concerning the global literature, organizational commitment revolves around the individual, the corporate, and job distinctiveness, which is aimed at discovering diverse critical relations between organizational effects and commitment.

## Employee Retention Strategies

Ramos (2019) formulated a taxonomy of five-talent retention approaches, which is linked to the criticality levels of the competencies of the employees. The study involved highlighting the criticalities of the competencies of the employees. These taxonomies consist of:

a. Competencies linked to internal knowledge of the companies such as leadership, technical skills or knowledge elements of the process and the product
b. Competencies related to the external experience of the consumer market or negotiation with the stakeholders. The following are some of the retention approaches:
    - Values and norms that are in line with the objectives of the company and a sense of belonging
    - Monetary and non-monetary incentives such as variable remuneration, salary increases, and career opportunities
    - Management of knowledge with manuals, information technology, specialized systems, and utilization of experience.

## Possible Reasons for Employee Retention

The numbers of organizational acquisitions and mergers have been increasing, which has always left employees displeased with their companies. During such cases, the employees are haunted by job security concerns. Employers also need to work to maintain their employees and prevent them from going to work for other companies or leave the organization and work for their competitors. Kumar & Mathimaran (2017) explained that employee development programs offered

by organizations are appropriate in retaining employees. Therefore, employee retention is essential in the long run since it is also a vital determiner of the success of the company. Maintaining the best employee ensures customer satisfaction, product sales, reporting authority, contented co-workers, effective progression planning, and deeply imbedded organizational knowledge and learning. Some of the possible reasons why employees leave a job include low wages or salaries, lack of reorganization, lack of growth and challenge, loss of trust in the manager or the supervisor for providing opportunities for self-growth, low job satisfaction, and lack of confidence in the senior management.

The current research, therefore, focuses on formulating predictive analysis in the real world. The study acknowledges that the Human Resource Department is the fundamental motivating aspect in contemporary business, providing the entrepreneurs with the appropriate motivating factor that steers the business success with the capital of course and its plan. More focus should also be directed to the provision of employee training since it enhances employee satisfaction. These measures are likely to minimize the cost of losing talented employees. Additionally, the strategy is expected to create more income for the business by enhancing employee output that mutually reduces costs, affects business output through improving talent management and employee retention, and filters the appropriate clients by sorting out their needs and fortifying their career paths. It is also essential to highlight the higher employee involvement with the plans of the company. These are achievable when the strategies are initiated in each Human Resource Department.

## Problem Statement

Organizations need to focus much on investing in an efficient training system. The problem that the study projects to focus on is the prediction of the necessity of training to assist in retaining the most talented staff. The key focus of the study is to predict the resigning employee to improve retention. It, therefore, means that the research focuses on the training aspect, an area that most companies have neglected, but is essential for employee satisfaction. To establish the most appropriate element for the organization and the specific talent that the company needs to maintain, the study formulated the approach to project employee training. The expectation involves learning about the management of people's expertise concerning organizational behaviours. The study also aims to develop original productive evaluation methods and applications in an actual scenario. It is also vital to manipulate the actual data at the expense of working with stored and real data that are likely to establish diversity as data scientists.

## Objectives of the Study

- **a.** To identify the talents which the company wants to develop and maintain to predict employee training.
- **b.** To determine the first-hand productive analysis techniques and implementation in a real-world scenario
- **c.** To highlight who is eligible to be retained in the company and have a successful career path
- **d.** To identify the employee retention rate in the organization under the study

## Research Questions

**a.** What talents does the company want to develop and maintain to predict employee training?
**b.** What are the immediate productive scrutiny methods and implementations in a practical scenario?
**c.** Who is eligible to be retained in the company and have a productive career path?
**d.** What factors are essential for developing the predictive model?

## Significance of the Study

The study is significant since it generates interests of employee retention from a broader perspective, and it is of crucial importance to those concerned with job performance. The results of the study will help provide the foundation for the work in the area of benefits package and development of competitive pay (formulating retention approaches that impact job performance). The research will also assist employers in implementing programs such as manager training approaches and flexible working arrangements to minimize turnover. The predictive model, which the investigation develops, will identify the need to provide training career, career counselling, career advancement, and development of employees within the workplace. The study will also contribute to the field of HR analytics as a way of refining the accurateness of the forecast of employee abrasion, and refining the solicitation for assisting project managers and HR professionals in improving the retaining rate of valued employees through establishing decision tree for identifying valuable employees and discovery factors that make them quit. Such strategies would save the employee turnover budget for the organization.

## Limitations

The time factor is the critical limitation of this study. The accuracy of the findings was also limited by the accuracy of the statistical tools employed in analysing the survey. Concerning the random forest algorithm used in the study, a proliferation of trees is expected to slow the procedure and therefore making it ineffectual for actual forecasts. Generally, the processes are fast to train but slow to formulate predictions once training is conducted. Other possible limitations include the enactment of the forecast system for employee retention because of less training data set. Also, the execution of the system for making decisions is limited due to the multifaceted growth of the random forest algorithm, accurateness of a decision consequence, inconsistent data, and inadequate admittance to the employee dataset.

## Literature review

Employee retention is an activity that starts at the top of every organization. The human resource department performs the function of hiring, sourcing, and retaining motivated employees. Retaining and sourcing for new employees necessitates recognized, focused, and comfortable procedures and policies, which makes retention a key management outcome. According to Begum and Brindha (2019), the HR department cannot minimize turnover. Leaders of companies need to set up significant and distinct positive change for the retention programs and processes within all organizational levels. The management also takes the responsibility of leading the employees

towards the performance of targets and goals after recruiting the right individuals. Therefore, the costs of high staff turnover are incredible since when an employee leaves an organization, the chances are that some are some substantial costs that are likely to arise (Begum & Brindha, 2019). These include enrolment costs, training costs, lost throughput costs, and lost sales costs.

According to Kumar & Mathimaran (2017), such actions impact businesses in various ways. For instance, since the reputation of the employees is at stake, the chances are that client might lose confidence transacting with a company that cannot retain its employees. Constant replacement of employees depicts poor management, lack of proper planning, and instability. Nevertheless, when a person plans to leave their work, often, it is the employee who is left bitter. They take the sensation along with them, combined with the abilities learned while working for the organization. On the same note, Kumar and Mathimaran (2017) added that such employees' opinions are often recurrent to forthcoming employers and their networks too.

Pradhan, Jena, & Pattnaik (2017) defined employee turnover as the departure or leak of intellectual capital from the organization. The study relied on voluntary turnover. The researchers established that the most resilient forecasters for voluntary turnover include tenure, age, pay, job gratification, and employee ideologies of equality and fairness. Other study outcomes of the same nature also recommended that demographic or personal variables, more so gender, age, marital status, ethnicity, and education were essential aspects in the prediction of voluntary employee turnover. Other elements that different studies focused on include job satisfaction, salary, growth potential, working conditions, and burnout.

According to Ribes, Touahri, & Perthame (2017), high turnover consists of different significant impacts on an organization. There are also challenges in replacing employees who are business domain experts or exhibit niche skill sets. These affect the productivity of the current employees and the on going activities within the organization or the company.

Ribes, Touahri & Perthame (2017) explained that organizations focus much on hiring employees. Organizations also focus on retention and development, and therefore the loss of an employee is significant damage to the company. Thus, the primary concern of managers should be to minimize the employee throughput to moderate the loss of the throughput. Ribes, Touahri & Perthame (2017) asserted that studies estimate that training and hiring of a replacement worker for the lost employee are costly, and costs about 50% of the yearly salary of the worker.

Other vital aspects that play a critical role as the projectors for turnovers include job involvement. According to Bhartiya, Jannu, Shukla & Chapaneri (2019), job participation is the aspect that specifies how much a worker is engaged in a specific job, and execute the situation from the start to end of various accountable roles. These enable an employee to feel the possession and inspire them to perform the work effectively. The key influences, which are detrimental in employee retention, provide a healthy work setting. Besides, if the reason for abrasion is established, there are possibilities of improving retention through involvement in discussion and planning with the employee.

More so, a study by Salunkhe (2018) suggested that the most resilient predictors of employee turnover are general job gratification, tenure, job performance, distinct demographic aspects (such as gender, ethnicity, marital status, age or experience), salary, geographical factors, job satisfaction, growth potential, and recognition. Salunkhe (2018) further explained that that push factor consists of employee mismatch with regard to the role performed or job requirement, work stress, unstable relational relationship with peers and the supervisor, the unacceptable work-life balance that makes an employee be unhappy, and subject the employee to resign from the organization. Deloitte (2017)

established that approximately 10, 000 HR and business leaders believe in the implementation of workforce analytics, and its significance in business performance.

Mitchell (2018) conducted another research on Indian companies. The study elucidated the aspects leading to attrition, such as low incentives, below expectation salary, and relationship with the superior. Others are lack of appreciation, skills recognition, and unsatisfied work culture. According to Archita (2017), the management gives rewards to the employees, which plays a critical role in retaining the employees. The interpersonal skills of the manager, therefore, play a crucial role in maintaining a valuable employee. The current research looked at the various factors which impact employee attrition in different organizations.

## CONDUCTING DATA MINING STRATEGY FOR PREDICTING ATTRITION

According to Alaskar, Crane & Alduailij (2019), the predictive model can be developed through diverse data mining methods. Data mining is, therefore, essential for human resource management since it helps in identifying different factors that impact employees for high attrition. According to Alao (2013), data mining is the process that relies on treasured ideologies, which can be obtained from an enormous dataset. One of these approaches includes the use of the Decision Tree as a critical tool for mining data. According to Alao (2013), Decision Trees are tree-shaped structures, which represent decision sets, and employ decision algorithms for mining data. The technique is also applicable in exploring the potential impacts of diverse inputs. Therefore, Alao (2013) explained that predictive analytics is primarily concerned with the prediction of the future by utilizing current and past data. The future forecast also relies on four things, which consist of causes of events, knowledge of past and present events, correct tools for predicting the future, and its accuracy through data mining techniques and statistical modelling**.** The developed model for projection of attrition requires checking for over and under-sampling. Bischl (2016) explained that in an under-sampling, there is a random replication of the majority category employed in training. Oversampling randomly replicates cases of the minority category. Among the research studies reviewed, some employed the use of Random Forest and SVM for predicting attrition of employees (Punnoose, 2016). The following are some of the ML algorithms researched on:

### Support Vector Mechanism

The approach is essential since it performs class categorization by establishing a division point for categorizing in the high dimensional space of a hyperplane. B-Gent (2006) explained that the larger the edge, the lower the inaccuracy of the classifier, and the error. According to Edouard (2017), the model is essential in solving linear and non-linear binary classification challenges, which make it be referred to as the maximum margin classifier.

### Decision Tree

According to Kotsiantis (2007), decision trees are trees applied in classifying cases by arranging them in line with the feature values. Every nodule in the decision tree is a representation of the feature in a case under classification, and every branch is a denotation of the value which the node assumes. The following figure is an illustration of the Decision tree.


**FIGURE 1**
**DECISION TREE**

Kotsiantis (2007) conducted a study of the decision tree and classified it as C4.5, which is an extension of Quinlan's earlier ID3 algorithm. Other studies utilize the C5 decision tree algorithm.

## Random Forest

According to Staniak & Biecek (2018), Random Forests emerge from the amalgamation of tree classifiers. Every tree relies on the approximations of an average vector, which is inspected separately. The algorithm is the key focus of this research since the study employs it in formulating a projection. Staniak & Biecek (2018) further explained that Random Forest exhibits characteristics of decision tree compared to the fitting for the training data set.

According to Staniak & Biecek (2018), Random Forests differ from standard trees since the latter mode in each of the algorithms is differentiated using the most appropriate approach among the categories of the predictors, which are randomly selected using the node (Punnoose, 2016**).**

## K-Nearest Neighbour (KNN)

According to Raschka (2017), the algorithm relies on condition-based learning. Characterization is also obtained from the fundamental significant part vote of the k-closest neighbours of every point. The two category phases using KNN relies on establishing the neigh boring data points and then decide on the category concerning their classes. Raschka (2017) further explained that an algorithm method is a non-generalizing approach because it maintains every training information in memory, and especially transformed as a KD or ball tree.

## Logistic Regression (GLM)

According to Rashcka (2017), logistic regression is a machine learning calculation for characterization. The estimation considers the possibilities that elucidate the potential outcomes of a separate trial, and they are represented using a logistic Regression.

## Factors for Deciding Valuable Employees

According to Ramos (2019), the critical focus of work engagement involves describing the degree to which employees engage in their activities. The job level and position determine employee engagement with work. Attridge (2009) explained that employees with Executive and Director Levels are majorly engaged in the work compared to the subordinate staff.

Attridge (2009) further explained that highly skilled workers are primarily involved in their work. Concerning this, the Human Resource Information System performs the critical role of making decisions for the appropriate management of the human resource. Attridge (2009) further explained Knowledge Discovery in Database (KDD) and Intelligent Decision Support System are applicable in HRIS since they improve the structured, more so unstructured and semi-structured HR in the decision-making process. The module also provides a total performance index, considering every criterion for the employee. According to Abdul-Kadar (2015), the process is multifaceted since there are a variety of rules for every criterion with a different priority. It is among the elements considered for identifying a valuable employee. Some skills for employees are predicted through Machine Learning and decision-making criteria with historical data.

Crossman's (2003) study suggested that employee performance is efficient if there is more job satisfaction. Therefore, job satisfaction is indirectly accountable for determining a valuable employee. Luthans (2008) also explained that the work, which is appropriate for the skill set of the employee, is essential in determining employee performance. According to Allan (2017), the skillset of the employee and the work relevance plays a crucial role in determining the employee's performance. Onsardi (2017) also explained that that one key factor that needs to be considered in assessing the performance of employee value is employee loyalty. For instance, it focuses on the number of years in which the employee has been working for the company. However, Onsardi (2017) relies on job satisfaction.

Ramos (2019) explained the need for identifying the personal employee's features among human resource departments. Other factors that the human resource departments need to consider include job attitudes and the work environment factors that are linked to employee turnover. Lack of such information would lead to the fragmentation of turnover control efforts, therefore the possibility of misguidance.

The accuracy of the earlier outlined algorithms is not resilient enough, and the elements of employee intent hardly concur in such models. One of the ways of addressing this involves augmenting the aptitude to conjecture on employee turnover and introduce new methods that can assist human resource departments to effectively identify the fundamental factors affecting employee turnover objective and predict employee turnover.

## Employee Retention Strategy

Ramos (2019) defined the term strategy as a planned and more formalized system of practices that are connected with a set of mission and values, overall vision. Most companies, especially large ones with fully established human resource departments who participate in intricate planning activities aimed at developing a unitary and cohesive strategy in handling employee retention, especially, the management of a human resource that majorly consists of retention as a key objective, are likely to formulate interventions and programs with an apparent reference to the general organizing principle.

It is also evident that that high turnover exhibits diverse harmful impacts on an organization. There are various challenges in replacing employees who have niche skill sets and are human domain experts. Replacement and acquiring new employees exhibit their costs, such as training costs, hiring costs, and so forth. New employees can also display their learning curves as a way of arriving at the same levels of business or technical expertise as a seasoned internal employee. Companies solve such challenges by employing machine learning as a way of predicting turnover, which provides the vision to take the appropriate actions.

### Summary and Justification of the Random Forest Method

The literature review, therefore, suggests that there exists a variety of technical and theoretical studies and research, which have been carried out to find attrition prediction. Nevertheless, there has been no critical research or study on the development of a tool that can execute automated decisions on classifying valuable and ordinary employees. There is also a lack of applications that display the final dashboard and indicates the retention aspects that HR managers need to take into consideration when retaining the critical employee. Thus, the Human Resource Management budget can be reduced in situations where the rate of retention is increased.

The Random Forest method is relevant to the current study because it does not require feature normalization or scaling. The technique also ensures easy measurement concerning the relative significance of every feature. The method also facilitates the handling of numerical and categorical features.

## METHODS

The methodology section provides the technical and theoretical insight of the research method employed in the study to establish the pattern, which links the employee elements in general, with the dependent variable (employee retention) considering factors such as gender, age, training, and so forth. The section aims to formulate a prediction model that is applicable in predicting an employee who is under retention risk by the use of the Random Forest Model.

### Research Strategy

The study surveyed the employees to identify the attributes and factors that are essential in predicting retention rates and improving retention of employees. The research conducted experiments with a machine-learning algorithm on the dataset, which increased the accuracy of the research outcomes.

### Data Set Overview

The dataset consists of the employment history of 613 employees with a single employer. The details captured in the data include:

- Employee ID
- Employee name
- Training count
- Date of birth
- Gender
- Marital status
- Engagement. date
- Contract type
- Leave entitlement
- Grade

- Employment title
- Employee's department
- Last action type taken against an employee by the employer with respect to their employment
- Action date
- The employment status
- Degree level achieved by an employee
- Degree major for the employees

The above variables are essential in helping an employer check the proportion of their staff that is leaving, the factors contributing to it, forecasting future terminations, and estimating the accuracy of these estimates. An overview of the data indicates the number of current active employees and the period of engagement with the company. The data spans up to 14 years from 15/01/2006 to 01/10/2019. During this period, a data summary indicates that 494 employees have left the company, 52 employees have had their contract types changed, and 73 new hires have occurred. The first action year in the data monitored by this project took place in 2007, while the last action year was 2019. The company conducts training on the employees with some employees receiving up to three pieces of training, while others are yet to attend any. The data also gives the contract types in the company, summary statistics on employee grades, and gender statistics. The company has employed 322 female workers and 297 males since 2016. As of 2019, the company had 125 active employees and 494 inactive, as shown in Table 1 below.

**Table 1**
**EMPLOYEE STATISTICS ON CONTRACT TYPE, MARITAL STATUS, TRAINING COUNT, LEAVE, GRADE, EMPLOYMENT TITLE, GENDER, AND DEPARTMENT OF WORK.**

| Contract type | | Mstatus | | Training count | | Leave.entitlement | |
|---|---|---|---|---|---|---|---|
| Graded | 195 | Divorced | 16 | Min. | 0.0000 | Min. | 0.00 |
| Lumpsum | 146 | Married | 285 | 1st Qu. | 0.0000 | 1st Qu. | 0.00 |
| DPE | 87 | Single | 214 | Median | 0.0000 | Median | 30.00 |
| Part Time | 79 | Widowed | 1 | Mean | 0.2149 | Mean | 19.62 |
| Service Order | 60 | | | 3rd Qu. | 0.0000 | 3rd Qu. | 30.00 |
| QCC | 29 | | | Max. | 3.0000 | Max. | 50.00 |
| Grade | | Employment. Title | | Gender | | Department | |
| Min. | 1.000 | Title 122 | 86 | F:322 | | Department 3 | 337 |
| 1st Qu. | 7.000 | Title 13 | 50 | M:297 | | Department 1 | 72 |
| Median | 7.000 | Title 172 | 20 | | | Department 7 | 59 |
| Mean | 8.042 | Title 82 | 15 | | | Department 6 | 34 |

Four hundred and ninety-three of the employees hired in the period under consideration hold a bachelor's degree, 40 have a master's degree, 27 a Ph.D., 25 a diploma, 17 primary school certification, 9 have a higher diploma, and ten others hold other certifications. The employer recruited 182 employees holding a business degree, 99 with a business and management degree, 27 with a Business Administration degree, 25 with a degree in IT, 15 with a degree in education, and 217 employees hold degrees and certification in various other disciplines. Table 2 below shows a summary of employee academic certification as well as degree majors.

**Table 2**
**EMPLOYEE ACADEMIC LEVEL AND DEGREE MAJOR**

| Degree Level | | Degree Major | |
|---|---|---|---|
| Bachelor | 493 | Business | 182 |
| Diploma | 25 | Business & Management | 99 |
| Higher Diploma | 9 | | |
| Higher School | 8 | Business Administration | 27 |
| Master | 40 | IT | 25 |
| PHD | 27 | Education | 15 |
| Primary School | 17 | (Other) | 217 |

## Data Collection Method

The research employed a survey in collecting primary data. Several case studies and research papers on factors affecting retention, employee attrition, and prediction of employee attrition comprised the secondary aspects of the data collection, as detailed in the literature review. The concepts theoretically studied concepts were practically applied to transform the business logic technologically. These involved the implementation of a Machine Learning Algorithm to predict employee retention and use a decision tree to classify the valuable employee from the ordinary one.

## Theoretical Framework

The significant retention aspects identified were employed as predictors of employee retention. The factors include the identification, gender, and training count, and marital status, type of contract, leave entitlement, grade, last action type, employment title, and engagement date. The retention elements were established to improve retention.

## Methodological Assumption

The methodological assumption for formulating the RF algorithm focuses on finding an accurate forecast that links employee characteristics with employee retention. The models employed include the RF and an advanced decision tree algorithm, as well as a linear model. Both the Decision tree and linear model confirm the consistency of results as displayed by the RF algorithm.

## Technical Framework

The fundamental concern of the technical framework consists of data processing and data selection. The research employs the Random Forest (RF) machine-learning algorithm for classification and regression. The data obtained is split into two segments: the training dataset and the testing data to avoid an instance of overfitting. The model is trained on 13 years of the employee training data and tested on the $14^{th}$ year data, which has been classified as the testing data. The predictors in the model include the last action type variable, training count, gender, and department. These variables are used to predict the number of employees in retention with the company in 2019. In essence, the model constructs decision trees using training employee data and outputs a prediction of the employees at the risk of retention in the given data set. Besides, the results from the RF algorithm are confirmed with the decision tree model as well as a linear model, which both show

consistent results that employee retention in the company is forecasted in a fair manner. The decision tree assisted in observing the occurrence of the prediction. The results from the random forest model support the research findings and objectives on employee retention rate with a forecast on the number of employees retained in 2019.

### Principle Component Analysis (PCA)

PCA has been used to reduce the dimensionality of the dataset to make it more interpretable while reducing information loss. The model under consideration uses data divided into two groups to avoid overfitting. Characteristics that have a high impact on employee retention are taken into model consideration while still preserving variability. PCA is a hearty statistical technique when dealing with such a large dataset.

## RESEARCH FINDINGS

The key focus is to highlight the findings obtained from the previous section and their application in progressing the research work—these range from data analysis that covers quantitative and qualitative research aspects. The section also highlights the outcome of the research in a manner that can be presented to the user. The assumptions were essential in completing the research work and establishing the eventual working analytical model.

### Employee Retention Rate

The research uses data from company X. The data ranged up to the last 14 years. The research employed the 2007-2018 data as the training data, while the 2019 data acted as the test data. The data provides information to help determine how retention occurs in the company across the years and in various categories such as departments and training. Figures 2 and 3 reveal how the data looks and gives a glimpse of a sample of the first six employees in the dataset.

| | ID | Employee.Name | Training.Count | D.O.B. | Gender | MStatus | Engagement..Date | Contract.Type | Grade |
|---|---|---|---|---|---|---|---|---|---|
| 1 | F10002 | Employee 2 | 0 | 29/01/1982 | F | Single | 01/01/2007 | Graded | 20 |
| 2 | F10003 | Employee 3 | 0 | 12/09/1963 | F | Married | 01/07/2006 | Graded | 12 |
| 3 | F10004 | Employee 4 | 0 | 18/12/1978 | F | Single | 12/11/2006 | Graded | 15 |
| 4 | F10005 | Employee 5 | 0 | 23/12/1979 | F | Single | 04/12/2006 | Graded | 14 |
| 5 | F10006 | Employee 6 | 0 | 29/01/1982 | F | Single | 12/08/2006 | Graded | 7 |
| 6 | F10008 | Employee 7 | 0 | 29/01/1982 | F | Single | 12/12/2006 | Graded | 7 |

**FIGURE 1**
**EMPLOYEE DATA OVERVIEW**

| | Department | Last.Action.Type | Action.Date | ActionYear | Status | Degree.Level | Degree.Major |
|---|---|---|---|---|---|---|---|
| 1 | Department 21 | Cancellation | 23/05/2007 | 2007 | Inactive | Bachelor | Business Administration |
| 2 | Department 21 | Cancellation | 01/07/2014 | 2014 | Inactive | Bachelor | Accounting |
| 3 | Department 5 | Cancellation | 05/04/2007 | 2007 | Inactive | Bachelor | IT |
| 4 | Department 3 | Cancellation | 29/04/2008 | 2008 | Inactive | Master | Adminstration |
| 5 | Department 19 | Cancellation | 10/03/2008 | 2008 | Inactive | Bachelor | Business Administration |
| 6 | Department 10 | Cancellation | 18/01/2009 | 2009 | Inactive | Bachelor | IT |

**FIGURE 2**
**EMPLOYEE DATA OVERVIEW**

The above data representation reveals that Company X had 613 employees in the dataset. The illustration also suggests that Company X had terminated 494 contracts since 2006, 52 contracts changed, and there were 73 new recruitments. The first contract was rendered inactive and terminated in 2007 but none in 2006. The initial action year was 2007, while the last one was in 2019. The active employees in 2019 were 125. It means that 494 employees were inactive. Of the employees hired up to 2019, 27 had Ph.D., 40 had master's degrees, 493 had bachelor's degrees, 25 diploma holders, nine higher diploma holders, 17 primary school certification, and ten others with other certifications. Out of the employees recruited since 2006, 182 had business degrees, 99 had a degree in management and business, 27 with a degree in business administration, 25 with degrees in IT, and 217 employees with degrees in other careers.

| | Active | Inactive | lost.by.company |
|---|---|---|---|
| 2007 | 0 | 29 | 100.00000 |
| 2008 | 0 | 74 | 100.00000 |
| 2009 | 0 | 103 | 100.00000 |
| 2010 | 0 | 48 | 100.00000 |
| 2011 | 0 | 31 | 100.00000 |
| 2012 | 0 | 42 | 100.00000 |
| 2013 | 1 | 34 | 97.14286 |
| 2014 | 0 | 16 | 100.00000 |
| 2015 | 0 | 14 | 100.00000 |
| 2016 | 2 | 36 | 94.73684 |
| 2017 | 4 | 22 | 84.61538 |
| 2018 | 73 | 28 | 27.72277 |
| 2019 | 45 | 17 | 27.41935 |

**FIGURE 3**
**ACTIVE AND INACTIVE LOST BY THE COMPANY**

**Percentage of Gender Leaving and Remaining with the Company**

The data revealed that 87.04902% of the staff left the organization since the first action date in 2007. Gender analysis suggested that 57 of the active employees were female (F), while 68 were male (M). More females left the company.

| | Active | Inactive | lost.by.company | remain |
|---|---|---|---|---|
| F | 57 | 265 | 82.29814 | 17.70186 |
| M | 68 | 229 | 77.10438 | 22.89562 |

**FIGURE** 5
**PERCENTAGE OF GENDER LEAVING AND REMAINING**

**Retention in Each Department**

The graph in figure 6 explains the retention trends in each department. From the resultant graph and table in figure 7, department 3 seems to be the most affected. The department is the most active with the current number of employees at 103 and 234, having left the same department.

## Strategies for Leaving Departments

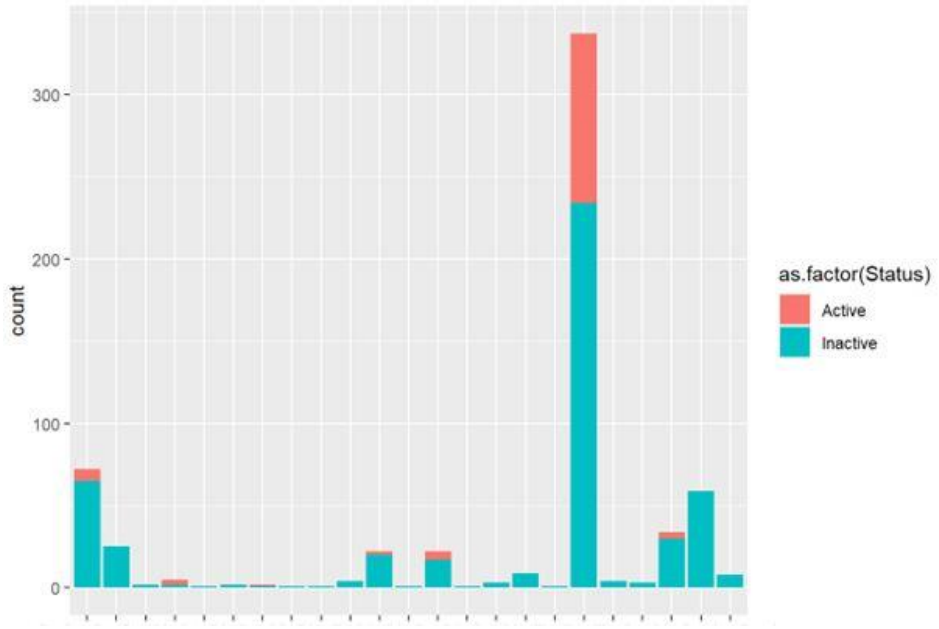


**FIGURE 4**
**RETENTION IN DEPARTMENTS, DEPARTMENT 3 HAS THE LONGEST BAR**

| | Active | Inactive | lost.by.company | remain |
|---|---|---|---|---|
| Department 1 | 7 | 65 | 90.27778 | 9.722222 |
| Department 10 | 0 | 25 | 100.00000 | 0.000000 |
| Department 11 | 0 | 2 | 100.00000 | 0.000000 |
| Department 12 | 3 | 2 | 40.00000 | 60.000000 |
| Department 14 | 0 | 1 | 100.00000 | 0.000000 |
| Department 15 | 0 | 2 | 100.00000 | 0.000000 |
| Department 16 | 1 | 1 | 50.00000 | 50.000000 |
| Department 17 | 0 | 1 | 100.00000 | 0.000000 |
| Department 18 | 0 | 1 | 100.00000 | 0.000000 |
| Department 19 | 0 | 4 | 100.00000 | 0.000000 |
| Department 2 | 2 | 20 | 90.90909 | 9.090909 |
| Department 20 | 0 | 1 | 100.00000 | 0.000000 |
| Department 21 | 5 | 17 | 77.27273 | 22.727273 |
| Department 22 | 0 | 1 | 100.00000 | 0.000000 |
| Department 23 | 0 | 3 | 100.00000 | 0.000000 |
| Department 24 | 0 | 9 | 100.00000 | 0.000000 |
| Department 25 | 0 | 1 | 100.00000 | 0.000000 |
| Department 3 | 103 | 234 | 69.43620 | 30.563798 |
| Department 4 | 0 | 4 | 100.00000 | 0.000000 |
| Department 5 | 0 | 3 | 100.00000 | 0.000000 |
| Department 6 | 4 | 30 | 88.23529 | 11.764706 |
| Department 7 | 0 | 59 | 100.00000 | 0.000000 |
| Department 9 | 0 | 8 | 100.00000 | 0.000000 |

**FIGURE 5**
**RETENTION IN DEPARTMENTS**

## Termination by Training Count

An analysis of retention based on training count to determine whether the company is losing highly trained people is depicted in figure 8 below. The resultant graph implies that most of the employees who are currently inactive did not attend any training. However, the company lost a small number of employees who had participated in training in 2018 and no any other year. When ordered by degree level, it is evident that most of the employees terminated between 2007 and 2018 hold a bachelor's degree. Recently, a mix of employees holding all degree levels were rendered inactive, as shown in figure 9. Besides, most of the employees who have left the company are of grades 6 and 7, as illustrated in figure 10.

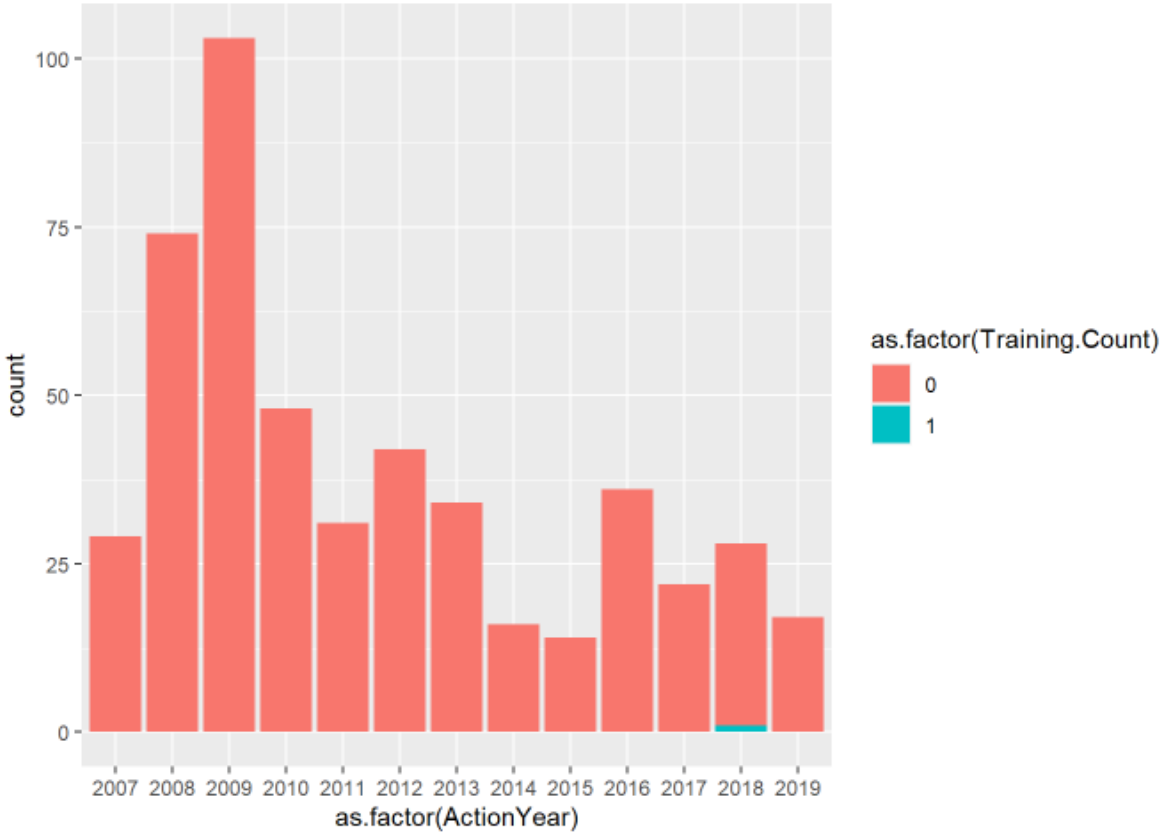


**FIGURE 6**
**RETENTION BY TRAINING COUNT**

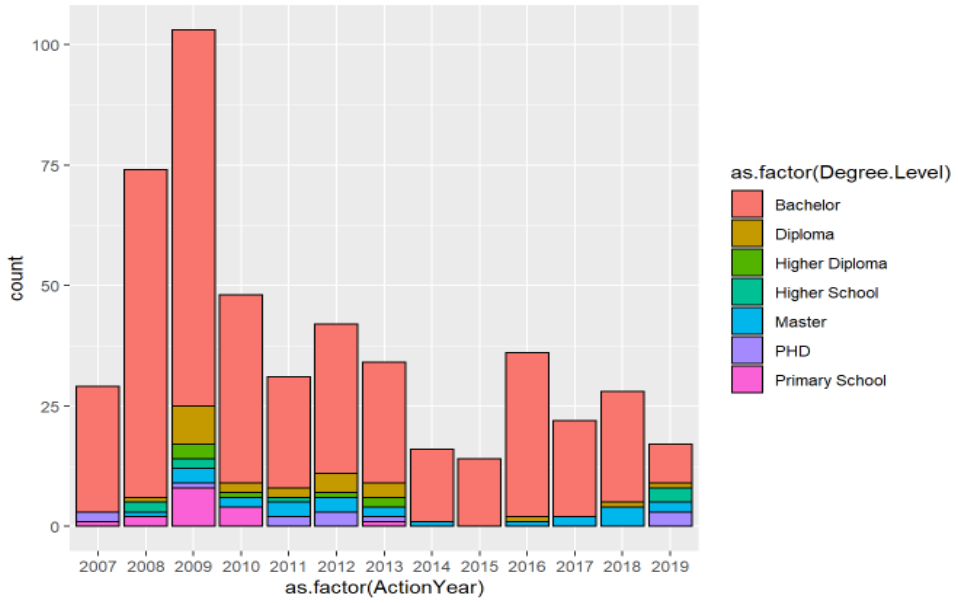


**FIGURE 7**
**RETENTION BY ACADEMIC QUALIFICATION**

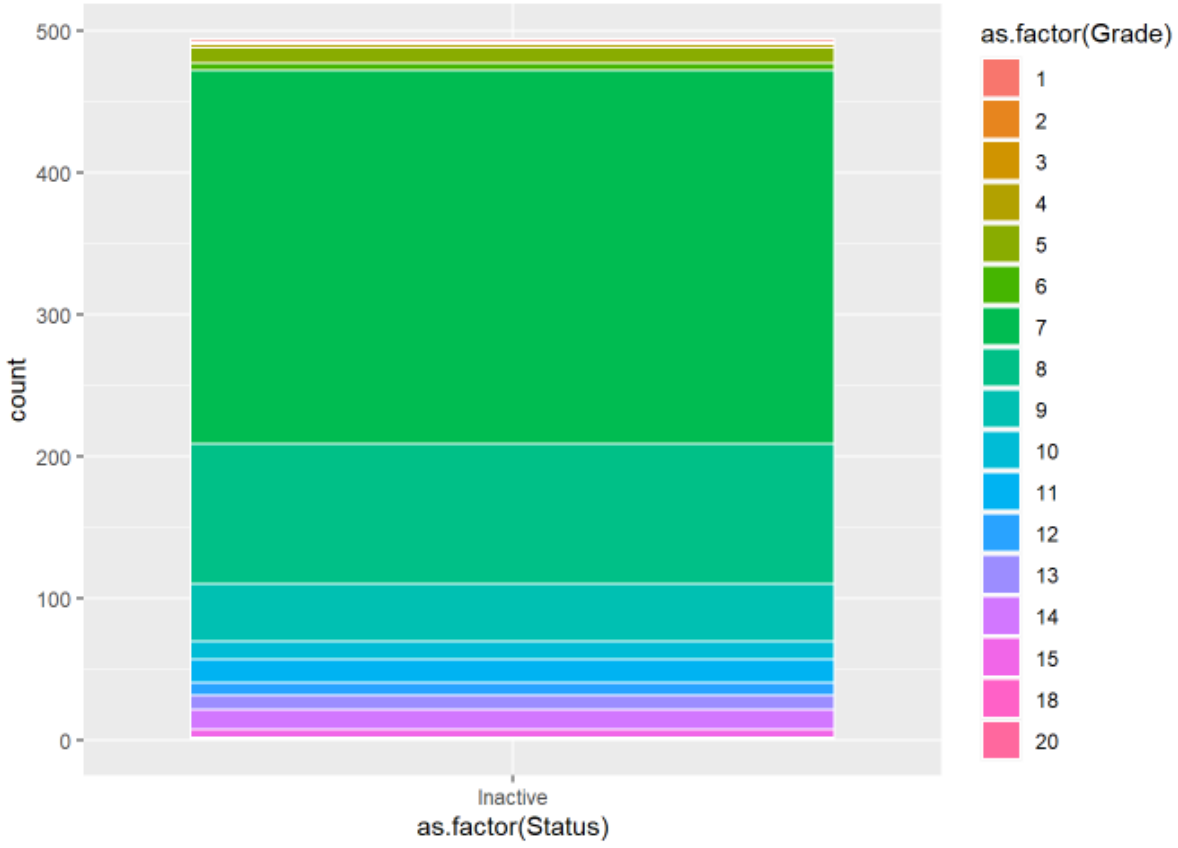


**FIGURE 8**
**RETENTION BY EMPLOYEE GRADE**

The above figures illustrate how termination occurred in Company X through training count. The statistics identify how Company X lost highly trained professionals. For instance, those who attended more pieces of training and with high academic qualifications. The majority of the inactive

employees had not attended pieces of training. Company X also lost some employees who attended pieces of training in 2007 and 2018 and had bachelor's degrees.

### Termination by Action Year and Contract Type

In 2007 and 2008, most of the employees in the graded category left the company as shown in figure 11. However, this trend changed in the next years as most lump sum employees left the company between 2008 and 2013. A huge number of DPE employees left the company in 2009. No lumpsum employee has again left the company since 2014. The recent years show that a moderate mix of QCC, part-time, and graded employees left the company. Most service order employees left the company in 2019.

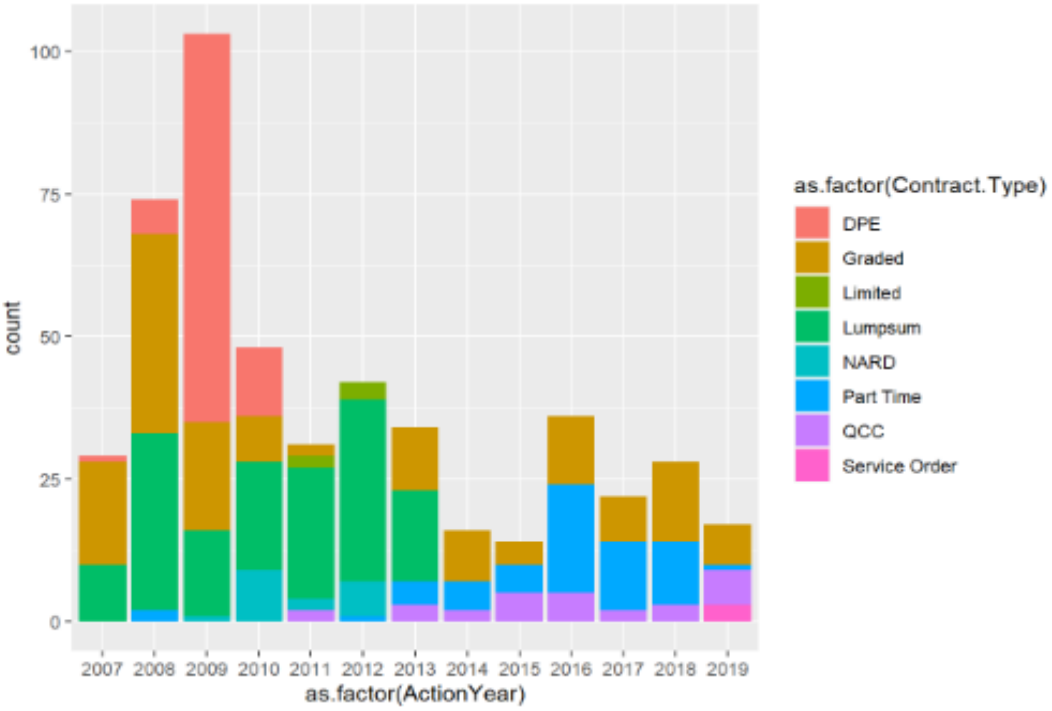


**FIGURE 9**
**TERMINATION BY ACTION YEAR AND CONTRACT TYPE.**

### Termination by Degree Major and Degree Level

From figure 12, it is difficult to explain whether the degree level with the respective degree major would cause the employee to leave the company. The retention rate on an employee's academic level and degree major is spread out. In essence, these variables are not very significant in determining retention rates.


**FIGURE 10**
**TERMINATION BY DEGREE MAJOR AND DEGREE LEVEL.**

## Model Building

To avoid over fitting, the data is partitioned into two categories. The model is trained on 13 years of data and tested on the 14th year. A variety of modeling algorithms are employed on the training and testing data. The algorithms show consistent results with the RF model. The variables selected to make the retention prediction and the data structure for prediction are shown in table 3 below.

**Table 3**
**VARIABLES SELECTED FOR PREDICTION.**

| No | Training | Gender | Leave | Grade | Action Year | Department | Status |
|---|---|---|---|---|---|---|---|
| 1 | 0 | F | 37 | 20 | 2007 | Department 21 | Inactive |
| 2 | 0 | F | 37 | 12 | 2014 | Department 21 | Inactive |
| 3 | 0 | F | 0 | 15 | 2007 | Department 5 | Inactive |
| 4 | 0 | F | 37 | 14 | 2008 | Department 3 | Inactive |
| 5 | 0 | F | 30 | 7 | 2008 | Department 19 | Inactive |
| 6 | 0 | F | 30 | 7 | 2009 | Department 10 | Inactive |

## Running Models

The models are run in the statistical software R. The software provides great functions and procedures to make forecasts. The research uses the R packages caret and rattle to predict the employees who might leave the companies in the future. The category predicted is whether a specific employee is Active or Inactive. Some of the helpful models for this result are the decision trees and linear models whose results are used to compare with the random forest. A plotted decision tree in figure 13 helps to see how the prediction is occurring. In essence, the model indicates that only 14% of the employees remain in Active status, while 86% of them have left the company, as shown in the figure below. An interesting revelation is that training count is shown to affect employee retention.

The random forest function in R is used to give a prediction of the employees at the risk of retention in the data set. There are 500 trees with two variables tried at each split. The default setting of m tries in the model yields the smallest OOB error. The OOB estimate error rate is 0.36%. The

area under the curve is 0.9927, and the confidence interval is 95%. The higher the area under the curve, the better. The random forest model shows that the most critical variables in predicting employee retention in the company in decreasing order are employee's Training Count, Action Year, employee grade, Leave Entitlement, Department, and lastly, their gender.

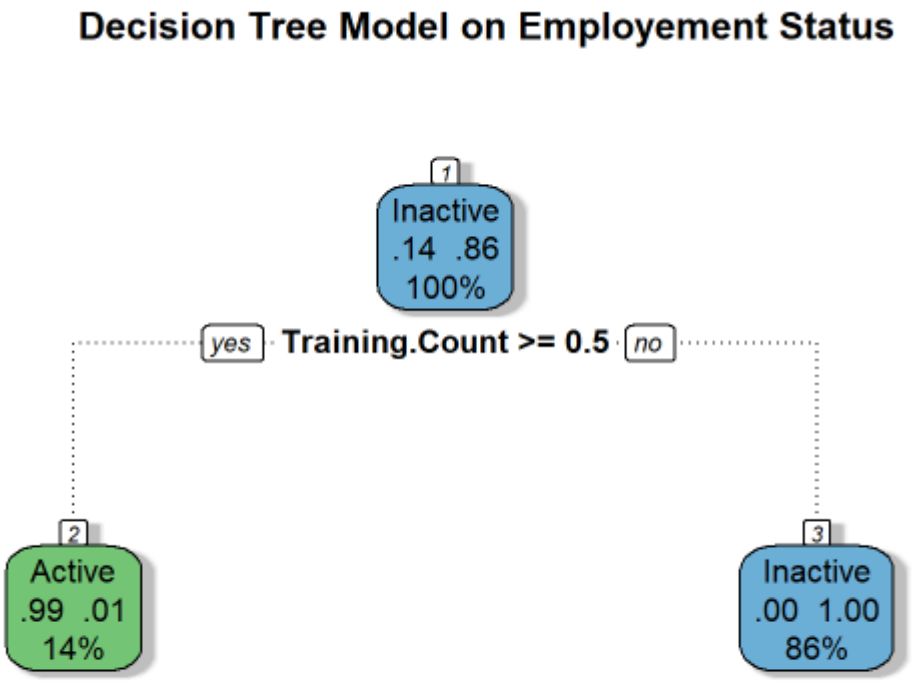


**FIGURE 11**
**A GENERATED DECISION TREE.**

```
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 0.36%
Confusion matrix:
         Active Inactive class.error
Active       79        1 0.012500000
Inactive      1      476 0.002096436
```

**FIGURE 12**
**RESULTS FROM THE RANDOM FOREST MODEL.**

The random forest model orders the importance of the variables in predicting, as shown in figure 15 below. The variables are arranged in ascending order of their importance. The most crucial variable in explaining retention is Training count, while gender is the least important.

| | Active | Inactive | MeanDecreaseAccuracy | MeanDecreaseGini |
|---|---|---|---|---|
| Gender | 2.33 | -1.35 | 0.56 | 0.35 |
| Grade | 10.59 | 0.17 | 10.01 | 3.44 |
| Department | 11.99 | -1.85 | 10.71 | 2.44 |
| Leave.Entitlement | 12.89 | 3.29 | 11.85 | 1.67 |
| ActionYear | 14.35 | 7.14 | 15.39 | 23.19 |
| Training.Count | 83.36 | 41.34 | 64.65 | 53.34 |

**FIGURE 13**
**IMPORTANCE OF VARIABLES IN THE PREDICTION.**

## DISCUSSION

The section explains the research findings, specifically to the objectives and how they were met with regard to the quantitative and qualitative approaches. The primary purposes of the research included identifying the talents which the company wants to develop and maintain to predict employee training; identify the first-hand productive analysis techniques and implementation in a real-world scenario; highlighting who is eligible to be retained in the company and have successful career path**;** and identifying the employee retention rate in the organization under study. The research achieved these objectives through discussing the employee retention rate, percentage of gender leaving and remaining, strategies for leaving departments, termination by training count, termination by action year and Contract Type, and termination by degree major and degree level.

Based on the research objectives, and the survey conducted among the HR professionals for Company X, the research found out the pattern that associates the employee characteristics in general and employee retention. The machine-learning algorithm employed in previous research was also used to enhance accuracy.

The dataset collected from the Random Forest analysis helped in predicting employee retention and provide the most accurate output. The study built the data model by partitioning the data by employing various algorithms to test and train data. Concerning the R models, they helped in providing excellent procedures and functions. These were essential in predicting the employees who might leave the organization in the future and predicted whether an employee is active or inactive in Company X. With the random forest, the research was able to provide a prediction of the employees who were at the risk of retention in the data set. The Random forest consisted of 500 trees with two variables tested at each split. The model employed a default setting of m tries in the model to obtain the smallest OOB error. The OOB estimate rate was 0.36%, while the area under the curve was 0.09927.

Similarly, the Random Forest model established that the confidence level was 95%. Concerning this, the higher the area under the curve, the better. The model also helped in highlighting the key variables that were essential in predicting employee retention in the company through decreasing order on the action year. These were employee training count, action year, employee grade, department, leave entitlement, and gender. The findings agree with those of Mitchell (2018), who found out that training was one of the factors leading to attrition. Other factors identified by Mitchell (2018) are low incentives, below expectation salary, and relationship with the superior, lack of appreciation, and unsatisfied work culture. Therefore, it is possible to evaluate the

models to check the possibilities of predicting employee retention. The study also employed error matrices as a way of tabulating the outcomes of the research with the predicted values.

## CONCLUSION

The company should strive to train more employees since those who have attended more training are retained. Besides, other variables like ensuring employees are entitled to leaves and good salary grades, would make them feel motivated and remain in the company. The departments with reduced retention rates need to be checked, and any dissatisfaction among employees cleared. The model highlighted that there are 79 active employees in Company X in 2019 and 1 inactive employees. The data indicated that there is a high number of individuals who left Company X between 2006 and 2019. Company X could spend a lot of money in training the employees, but does not gain from such investment due to the large number of individuals who leave the company to work for other organizations. However, a keen look reveals that most employees who leave the company are those who have never attended any training. The HR department of Company X can utilize the obtained data to formulate a data-driven employee retention decision about the outcome of the model since the most critical variables for prediction are highlighted. The revelation that most trained employees remain with the company indicates that the retention strategies in the company are good. The RF model is applicable in the HR department to predict the overall state in the HRM process, which include employee retention, hiring new employees, attrition, the amount spent on training and developing new employees.

## REFERENCES


Masum, A.K., Beh, L.S., Azad, A.K., & Hoque, K., (2015). Intelligent human resource information system, *15*(1), 121-130.

Alao, D., Adeyemo, & A.B., (2013). Analyzing employee attrition using decision tree algorithms. computing, information systems & development informatics, *4*(1), 17-28.

Alaskar, L., Crane, M., & Alduailij, M. (2019). Employee turnover prediction using machine learning. *International Conference on Computing*, 301-316.

Crossman, A., & Zaki, B.A. (2003). Job satisfaction and employee performance of Lebanese banking staff. Journal of Managerial Psychology, *18*(4), 368-376.

Banerjee, A., Gosh. R. K., & Gosh, M. (2017). A study on the factors influencing the rate of attrition in it sector: based on indian scenario. Pacific Business Review International, *9*(7), 1-13.

Attridge, M. (2009). Measuring and managing employee work engagement: a review of research and business literature. Journal of Workplace Behavioral Health, *24*(4), 383-398.

Gent B, Coussement. K., & Poel D.V.D. (2006). Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques.

Bhartiya, N., Jannu, S., Shukla, P., & Chapaneri, R. (2019). Employee attrition prediction using classification models. In *2019 IEEE 5th International Conference for Convergence in Technology (I2CT)*, 1-6.

Ribes E, Karim, T., & Benoit P. (2017). Employee turnover prediction and retention policies design: A case study, *US: Cornell University*.

Luthans, F., Steven M.N. Bruce J.A., & James B.A. (2008). The mediating role of psychological capital in the supportive organizational climate: Employee performance relationship. Journal of Organizational Behavior, *29*(2), 219-238.

Kotsiantis, S.B. (2007). Supervised machine learning: A review of classification techniques. Informatica, *31*(1), 249-268.

Kumar, A. A., & Mathimaran, K. B. (2017). Employee retention strategies: An empirical research. *Global Journal of Management and Business Research, 17(1)*.

Hoffman, M., & Tadelis, S. (2018). People management skills, employee attrition, and manager rewards: An Empirical Analysis. National Bureau of Economic Research.

Nafeesa Begum, A., & Brindha, G. (2019). Emerging trends of it industry policies for ensuring women employee retention. *Indian Journal of Public Health Research & Development*, *10*(3).

Onsardi, A.M., & Abdullah, T. (2017). The effect of compensation, empowerment, and job satisfaction on employee loyalty. *International Journal of Scientific Research and Management, 5*(2), 7590-7599.

Pradhan, R. K., Jena, L. K., & Pattnaik, R. (2017). Employee Retention Strategies in Service Industries: Opportunities and Challenges. *Employees and Employers in Service Organizations*, 53-70.

Ramos, P.C. (2019). Employee retention strategies for executive operation leaders in an academic nursing environment, Walden University.

Raschka, S, & Mirjalili V. (2017). *Python machine learning*. S.l. Packt Publishing Ltd.

Ribes, E., & Touahri, K., & Perthame, B. (2017). Employee turnover prediction and retention policies design: a case study, *Cornell University*.

Punnoose, P., & Ajit, P. (2016). Prediction of employee turnover in organizations using machine learning algorithms. *(IJARAI) International Journal of Advanced Research in Artificial Intelligence, 5*(9), 22-26.

Salunkhe, T.P. (2018). Improving employee retention by predicting employee attrition using machine learning techniques, *Dublin Business School*.

Staniak, M., & Biecek, P. (2018). Explanations of model predictions with live and break down packages. *arXiv preprint arXiv:1804.01955*.