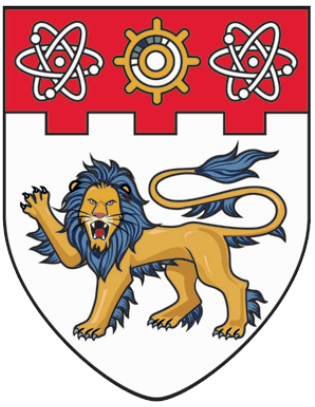


NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

# BC2406 Analytics I: Visual & Predictive Techniques

## Academic Year 19/20 Semester 1 Project Report


| | | |
|---|---|---|
| Seminar Group: | 4 | |
| Team Number: | 3 | |
| Team Members: | Agnes Ng Yi Rong | (U1810572B) |
| | Ernest Chua Hui Sheng | (U1810756H) |
| | Jalvin Nai Guang Jun | (U1810925F) |
| | Jaslyn Tan Chiew Yee | (U1810633J) |
| | Tan Jing Yu, Denis | (U1810101D) |
| Professor: | Hyeokkoo Eric Kwon | |
| Submission Date: | 13 November 2019 | |

# TABLE OF CONTENTS

## Executive Summary

Employee attrition, defined as the reduction of staff due to voluntary or involuntary reasons, has always been a problem for companies to deal with. High opportunity costs, amongst other business problems, arise out of employee turnover and prove detrimental for companies to deal with. Thus, this report attempts to investigate what goes wrong in the hiring process to cause high employee attrition and address the issue with the use of modern analytics to supplement the hiring process.

With the advent of technology, the team is hopeful in delivering positive results to help companies identify suitable candidates based on their skill sets and experience and match them with relevant job openings to improve companies' efficiency, productivity and accuracy in job matching. Thus, the analytic problem in this report is to identify the significant variables and generating a model capable of determining employee attrition.

Based on a fictional data set collated by IBM on what is assumed to be a pharmaceutical company, the team has conducted a thorough analysis on the dataset to gain deep insights into the reasons behind employee turnover. The data set was first cleaned and prepared for modelling. Singling out variables that would be available to a company at the initial hiring stage and applying logistic regression modelling and the team has identified significant variables that influence an employee's attrition status within the company as well as reliably predict turnover in a company. Utilising Classification and Regression Tree (CART) modelling, the team has identified variables to classify employees into correct job openings which reduces the mismatch of skills and education with job openings. Another logistic regression model was employed on a subset of data which included only current employees and terminated ex-employees to derive information on why employees were terminated and from these findings, the company could avoid hiring such individuals in the future. Due to the high costs of employee turnover, this report shall also investigate the relevant variables in retaining employees.

Applying our findings and the analytic solutions presented in the report, the team has also proposed business recommendations for the company in view of solving its business problems. These include the addition of a database which makes use of the logistic model prepared to increase efficiency in the hiring process and applying the CART model to allow a better match of skill set and experience to job openings and proposal of changes to the company and organisation culture to retain employees based on results from the logistic regression model.

Lastly, the report discusses and acknowledges the limitations of our solutions and research.

# 1 INTRODUCTION

Today, companies rely heavily on human resources. As a result, the hiring process is critical to identifying the right person for the right job. The traditional recruitment process, however, is not efficient in helping companies identify the most suitable candidate. Namely, there are a few problems with the traditional recruitment process which can be mitigated with appropriate analytics solutions.

First, companies have to manually evaluate a high volume of applications daily to shortlist potential candidates. This is done by assessing past experience, candidate skill-set and qualities for the relevant position. This process is labor-intensive, time-consuming and inefficient. Hence, there is room to improve the effectiveness of the current process especially for roles with high turnover rates such as sales, where attrition can hurt a company.

A modern approach would be required to identify potential candidates with suitable qualities and appropriate prior experiences. This would effectively help a company reduce cost and time spent sieving out individual candidates from a large pool of applicants. Also, the modern recruitment approach can make use of social media such as analysing an applicant's LinkedIn profile or social media profiles like Facebook or Instagram to screen potential candidates. It provides a look into someone's personality and identifies any possible red flags before any resources and efforts are spent.

This is where the team believes analytics could be used to supplement the hiring process. Given the candidates' data, analytics could give a forecast for the qualities required for a certain job role as well as giving a reliable estimate of the tenure of an employee. This translates to more accurate results as opposed to the traditional hiring process. Since analytics could provide valuable insights on the viability of a candidate, a better allocation of resources can also be arranged to ensure resources are channeled to the "right" employees; employees who plan to stay for the long run.

# 2 BUSINESS ANALYSIS

## 2.1 Industry Background

Every company has different departments that require slightly different hiring processes. Hence, the most prominent problem within each department has to be identified so solutions could be provided to address each problem. In our report, the team will be analysing the pharmaceutical industry and propose solutions to improve the current hiring process of companies within the industry.

The pharmaceutical industry is one of the most important industries in driving medical progress. With rapid advancements in science and technology, the

research-focused industry is set to advance and provide more innovative solutions such as personalized medicines for patients all around the world.

Within the high technology sector, the pharmaceutical industry has the highest added-value per person employed **(EFPIA, 2018)**. This is much higher than other high technology and manufacturing industries. One main reason why employees in the industry are so valuable is the wealth of knowledge they bring to the company. The industry requires its employees to be highly educated. Research positions often require a master's or doctoral degree **(see Appendix A-1)**. The industry also came out on top in terms of the R&D investment to net sales ratio **(see Appendix A-2)**. On a global level, the pharmaceutical industry is projected to be valued above US$1.5 trillion by the year 2023. **(Aspa, J., 2019)**. This means that pharmaceutical companies can only draw talent from a limited pool of people. With such immense competition and demand for capable employees, it is crucial that companies have efficient processes to identify potential employees to hire.

## 2.2 Existing Business Problems

### 2.2.1 Education and Skills Mismatch in the Research and Development (R&D) Department

The R&D department is often the largest department of a pharmaceutical company. It usually contains the largest amount of employees as R&D is what drives the multi-billion dollar industry and helps companies profit. As such, companies inject large amounts of investment in R&D activities.

However, the R&D department requires a specific skill set and education. To conduct research and development on products, companies require employees to have extensive experience and studies in the medical, chemical or biological fields. Moreover, the pharmaceutical industry, being knowledge-intensive, often requires senior employees to be well-versed in both technical and managerial skills in order to manage complex situations **(N. Guliwe, personal communication, 2010)**. These managerial positions are extremely crucial as they involve strategic decision making which ultimately affects the performance of the company. It is not easy to find candidates with this combination of skill sets **(Van Zyl, 2009)**. This has resulted in an extremely competitive hiring process in the industry where talents are scarce and staff turnover is high **(Sanofi-Aventis, 2010)**.

New medicines delivered to the market requires extensive preclinical development and clinical trials before being approved. It takes on average 12-13 years for a new medicine to be delivered to the market from the first time it was conceptualized **(see Appendix A-3)**. All these require the continued efforts of the R&D team over the years. These research scientists are pivotal to the creation and discovery of new knowledge which eventually leads to the innovation process.

Due to the criticality and the scarce supply of suitable talent, companies frequently deal with the problem that job vacancies remain open for an extended period of time. Until these job vacancies have been filled up, the companies have to absorb the costs and potential revenue loss associated with the vacancies **(B. Letsoalo, personal communication, 2011)**.

Human capital in the pharmaceutical industry is salient and a persistent mismatch of skills and education could deprive companies of the productivity and efficiency needed to thrive in this competitive and demanding industry.

While the focus has been on the research department in the preceding section, two other departments are equally crucial to the operations of any pharmaceutical industry. They are the human resource (HR) department and the sales department.

### 2.2.2 Rising Burden on Human Resource Department due to Outdated Hiring Practices

Apart from managing the employee recruitment process, the HR department encompasses all aspects of people management, communication and is pivotal in building a positive culture. Hence, if many people are resigning, it may suggest that the current hiring process is inefficient and extremely tedious for HR employees.

As mentioned earlier, the industry constantly struggles to attract top science graduates and postgraduates. These talented individuals are often demanded by other blue-chip companies in various sectors as well **(Mohan, A. C, 2015)**. A survey found that 57% of European pharma companies find it difficult to drive innovation as they are not able to hire the right people with the right skills **(Megget, K., 2018)**. The recruitment process in the pharmaceutical industry has become highly competitive and the HR department plays an ever-increasingly important role in talent acquisition and management. Essentially, the HR department forms the platform on which the entire company would build on by hiring suitable employees.

However, based on traditional hiring methods, the hiring process has proven to be inefficient and taxing on resources. Huge amounts of time are expected to be spent screening every single applicant for a job and in an industry where employees are expected to have certain skills and qualifications, sieving out these ideal candidates can take a long time. Moreover, due to the difficulty in selecting employees and the huge competition for the supply of employees, mistakes in the hiring process could turn into a vicious cycle for the HR department, where there is an endless search for new employees. This could waste valuable resources and time which could have been better invested in generating productivity for the company. Thus, it is important for pharmaceutical companies to accurately identify suitable personnel for their departments.

### 2.2.3 High Turnover Volatility in the Sales Department

Due to the scarce supply of graduates and the high demand for competent employees in the competitive industry, many employees leave for better opportunities or are poached. All in all, it has generated a volatile turnover rate for the pharmaceutical industry - even more so for the sales department where traditionally, sales departments have had high attrition rates. Salespersons leave companies easily due to the commission-based nature of their jobs and the competitiveness of the industry.

In the global pharmaceutical industry, its sales force has always been a key driver of sales and revenue growth **(Quantzig, 2018)**. The R&D department creates lucrative products and Sales have to market these products for profits. Sales Force Effectiveness (SFE) is a concept commonly used in this industry. It aims to boost organizational revenue by leveraging customer acquisition strategies, improving marketing plans and targeting highly profitable customers **(Zoltners et al., 2010)**. A research conducted by McKinsey & Company suggested that high performing employees in management roles increase profits by 49% while high performing salespeople are able to increase profits by 67% **(Hejase, H. J., & Dirani, A. E., 2016)**.

Hence, to maximise the companies' profit outlook, pharmaceutical companies need to reevaluate their hiring and retention strategies for the sales department which faces more volatility compared to the other two departments.

### 2.2.4 Challenge in Assessing Talent for Cultural Fit during the Hiring Process

The term "cultural fit" is more than just a buzzword in the corporate world. According to Harvard Business Review, corporate culture is defined as the "glue that holds an organization together" and cultural fit is the likelihood that an employee will adapt to the core beliefs, attitudes and behaviors of the organization **(Bouton, K., 2015)**.

In the pharmaceutical industry where knowledge and experience are highly valued, it is convenient to believe that candidates who possess the highest education level or years of experience in the industry have the highest chances of integrating well into the company culture **(Severi, T., & Harap, D., 2017)**. However, this is often far from the truth. Marilyn Nyman, an organizational consultant, made this statement "Flawless execution – the precise area where future leaders will be expected to excel – is impossible if you bungle the culture challenge." **(PharamExec, 2010)**. The culture challenge is an obstacle for both startups and big-name pharmaceutical companies. This is due to the fact that workplace culture changes from place to place and companies constantly struggle to integrate employees into the company's culture.

A report by PWC found out that close to a third of all new employees leave the company within the first year of hiring, often due to cultural mismatch **(PWC, 2012)**. Turnover is undesirable in any industry. However, the effects of turnover are usually exacerbated in the pharmaceutical industry where the success of a company is heavily dependent on its ability to remain agile and responsive to any changes in the healthcare landscape. The report also mentioned that it cost the company 50-150% of an employee's annual salary to replace the position. Worse still, the fact that most employees do not work at optimal productivity during their initial year of hire provides an exceptionally low return on investment for companies that experience a significant amount of employees that leave within the first year **(PWC, 2012)**. By accurately assessing each potential employee before hiring, this will greatly decrease hiring costs for the company and at the same time, prevent any turnover-related obstacles which could hinder the company's progress.

### 2.2.5 Loss of Resources Due to Turnover of Employees

Employee attrition, especially in the form of voluntary resignation is costly to the company. As a cumulation of the above business problems, the failure to identify suitable candidates and a lack of ability to retain employees would lead to higher turnover rates. The company would have lost its human capital investments in the form of resources spent on training, workshops and other forms of employee enhancement.

When employees leave the company, they take with them their skills cultivated in their time at the company which cannot be reclaimed by the company. Companies would then have to spend more time and resources to hire a new employee, which means they would have to go through the entire hiring process again, which is costly. There would also be an opportunity cost associated with this since new employees would require time to acclimatize to the company's culture and the workflow before being able to perform to the best of their ability **(Surji, Kemal, 2013)**. Ultimately, employee attrition impedes on a company's productivity and efficiency. It is also disruptive to the organisational culture of a company when turnover rates are too high. Thus, it is imperative for companies to minimize their employee turnover via efficient hiring processes.

## 2.3 Business Opportunities

Job matching has become one of the vital factors affecting job productivity. According to the Organisation for Economic Co-operation and Development (OECD), 1 in 3 people among the global labour force had the wrong skills needed for their particular jobs **(The global skills mismatch, 2019)**. This shows the severity of a job mismatch and how pertinent it is in today's job market. Job mismatch refers to having discrepancies in terms of the educational field and vocation. Not only does the mismatch of skills affect the company, but it also affects the entire economy.

From the economy's perspective, this problem translates into a loss of resources and human capital which could bring about dire ramifications on overall productivity. The estimated cost as a result of this mismatch of skills is an annual GDP loss of US$5 trillion. Coupled with the rapid changes in technology, the level of skill required for a certain job is constantly changing which only makes it increasingly difficult to find a suitable person for the job. Instead of acquiring more skilled labour, companies need to better allocate their employees to maximise efficiency and productivity. Being able to improve the employee allocation process would translate to a huge amount of money and time saved for the company.

### 2.4 Expected Business Outcome

With the main business problem identified in the previous sections, the goal of the company would be to reduce the attrition level of its employees while retaining the current employees within the company. At the same time, the company would be able to better profile their employees based on the qualities which current, non-terminated employees possess, aiding in the screening process at the pre-interview stage. This would ensure that the company's resources are channeled towards employees who would stay in the company, which would mitigate the problem of loss of resources due to the turnover of employees. Specific to each department, the identified problems should ultimately be targeted and resolved to ensure the efficiency of the company.

The mismatch between education level and job role in the R&D department should be mitigated in order to ensure that employees in their job role feel sufficiently challenged and stimulated such that they attain a higher level of satisfaction while doing their jobs.

With respect to the high turnover in the Sales Department, the eventual goal of the company would be to identify factors in which other companies may be offering or variables which are discouraging the sales employees to stay. Thereafter, they would be able to act on these variables, increasing the chance of employee retention as well as a better working environment within the company for its employees.

Pertaining to the problem of a rising burden on HR, this would naturally be solved granted the two aforementioned problems in the R&D department and the Sales department since employees are less likely to leave which also eradicates the need for HR to hire a replacement for them.

Lastly, being able to filter applicants who are able to adapt to the company's culture would alleviate the issue of challenges in assessing applicants for cultural fit during the hiring process.

## 3 DATA SET

To demonstrate how the HR department can leverage external data and non-traditional methods to improve accuracy in the matching process, our team will be using and carefully analysing a fictional data set created by IBM data scientists with more rows added by another user. The data set consists of 23532 rows and 37 columns. It is suitable for analyzing the recruitment process and attrition problems faced by the pharmaceutical industry.

There are three departments in the dataset – Human Resources, Research & Development and Sales. The team would focus on the attrition variable from this dataset as the outcome. The team's methodology includes splitting the data set into 2 subsets before further analysis – IBM1 would contain current employees and former employees who voluntarily resigned while IBM2 would contain current employees and those who were terminated.

IBM1 could be used to garner insight on why employees chose to leave the company since individuals on the data set are wanted by the company and they are either current employees or they chose to leave. On the other hand, IBM2 could group similar profiles of employees who got terminated and from these findings, the company could refrain from hiring similar candidates in the future.

To go one step further, we have also separated the variables presented to us into two main groups : Group (1) variables are those that the company can obtain from the potential candidates during the initial hiring stage whereas Group (2) variables are metrics the company can only obtain from current employees which would be used to project onto the profile of future employees. **(see Appendix B).**

Group (1) variables would be used to filter employees based on demographic data obtained. This would allow the company to craft a reasonably accurate picture of what a loyal employee will look like which they can use to refer to during hiring. On the other hand, Group (2) variables can be reasonably controlled or influenced by the company for current employees. Hence, the company can alter and change its company policies to suit the qualities in which its employees desire.

## 4 ANALYTICS APPROACH

### 4.1 Analytics Problems

To determine the correct employee for the correct job, we would need to map significant variables to draw a conjecture of which variables lead to the high attrition rate. In this case, the "correct" employee would be one that is not terminated nor would he/she eventually voluntarily resign. This would require the construction of an

accurate predictive model to predict the potential employees during the selection stage based on certain criteria.

However, with only limited data that can be collected during the hiring stage, it might not be sufficient for a company to extrapolate on how good a fit a potential employee may be for the relevant job openings. Having said that, external data would need to be leveraged upon. While there was data to collect in the initial hiring stage, we concluded that some essential data is inaccessible to predict whether a potential employee would stay for the long term or be the right fit for the job. Hence, we would need to rely on the data from current employees to supplement and provide a projection onto what an employee's profile needs to be like in order to be able to be a good fit for the company. This would be similar to providing a checklist to match the employee's characteristics to the requirements of the company.

### 4.2 Analytics Solutions

#### 4.2.1 Analytics Solution 1

To aid the company in identifying the most suitable candidate who would fit in the company, a logistic regression model could be utilised. A model, with the help of the data set of current and past employees, could undergo supervised learning to predict the attrition status of a future employee as well as allow the company to derive significant variables which a candidate should possess in order to hire him and ensure lower employee turnover rates.

Also, given the data set of those employees terminated, the team could proceed with analysing the data and garner insights and trends for why these individuals were terminated. With these insights, the company could avoid candidates with such traits to save precious resources which could have been expended on the wrong employees.

#### 4.2.2 Analytics Solution 2

A CART model could be employed to ensure employees are placed in the right job openings. The decision tree could allocate a candidate, given his credentials, qualifications, experience and skillsets into a suitable department or job role. Given the data set of current employees, this could be done easily with high accuracy, assuming the company is content with the placement of its current employees.

#### 4.2.3 Analytics Solution 3

Next, the team has deemed it equally important to retain employees after identifying and hiring suitable candidates. As such, we could make use of the data set and derive significant variables from a logistic regression model. These variables should

then be the focus of the company - efforts would have to be made on these areas which have been proven by the model to be crucial in retaining employees.

## 5 DATA PREPARATION AND UNDERSTANDING

### 5.1 Data Cleaning

Before the team can proceed with utilising the data set and conducting analysis to derive insights and trends, data cleaning would have to be performed.

| What went wrong? | What was corrected? |
|---|---|
| Inappropriate values | "", "missing", "na", "Test", "Test 456", "TESTING", "?????", "TEST"<br>were changed to NAs. |
| Duplicates | Removed duplicate rows (rows where all values were exactly the same). |
| Redundant columns | EmployeeCount, Over18, StandardHours - constant values with no predictive value.<br>ApplicationID and EmployeeID - Unique identifiers with no predictive value. |
| Cleaning specific data | Perhaps due to data entry errors, certain rows had values that made no sense.<br>e.g. “1296” in “Department”’. Removed these entries. |
| Data type conversion | Converted relevant columns to factor and integer data types. |
| Data validity check | Create a variable, AgeStartedWorking (TotalWorkingYears - Age). This allowed us to find logical errors where AgeStartedWorking < 0, which was impossible. Removed such entries.<br>Check for TotalWorkingYears less than YearsAtCompany which was impossible. Removed such entries. |
| Remove all rows which have at least 1 NA | There were 273 NAs out of 23432 rows by 30 columns, which we deem to be insignificant. |

The final cleaned data set is exported in a CSV format and used subsequently for our analysis in the later parts of the project.

### 5.2 Feature Engineering

Feature engineering enables us to generate deeper insights and create better machine learning models by decomposing or aggregating features to create new features. Through feature engineering, we are able to integrate domain knowledge and expertise into our model construction process resulting in better model performance.

| Variable | How it was derived | Rationale |
|---|---|---|
| PriorYearsOfExperience | TotalWorkingYears less YearsAtCompany | This acts as a proxy to gauge an employee's level of experience before joining the company which provided a more intuitive explanation regarding the difference in attrition outcomes despite the age differences. |
| AverageTenure | PriorYearsOfExperience divided by NumCompaniesWorked | This was to investigate whether employees had a "job-hopping" culture. Furthermore, this would suggest to the company how loyal the employees would be. Hence, a more efficient resource allocation process can be planned to reward employees who are likely to stay for the long term and even avoid hiring candidates who have a higher propensity to leave. |

## 6 EXPLORATORY DATA ANALYSIS

Firstly, the proportion of terminated employees is too small to result in any significant modeling or to have any predictive value. Also, the proportion of current employees is large to the extent that we felt it might skew results. In order to solve this, we split the data into IBM1 and IBM2 where IBM1 focuses on current employees and those who have voluntarily resigned. IBM1 is the main focus of our analysis - it is what the company 'wants'. It includes current employees - since they are still working, IBM deems them the right person for the right job, as well as those who resigned. Basically, those who resigned are the right people according to IBM but the employees themselves feel a need to leave the company, be it for unsatisfactory conditions or better opportunities elsewhere. Therefore, we are using IBM1 to find

out the variables which would indicate/suggest how to keep the right person for the right job. IBM2 has current employees and those who were terminated. We are using IBM2 to find out what variables cause termination – which is the employees whom the company deems as a mismatch (the wrong person for the wrong job).

### 6.1 Univariate Analysis

We will be looking at key demographic aspects of the entire company's employees to gain a better understanding of their profiles.

| **Variable** | **Main Findings** | **Appendix Reference** |
|---|---|---|
| Attrition | Majority are current employees | Appendix C-1 |
| Age | Majority from age range of 30-40 years old | Appendix C-2 |
| Department | Majority of the employees belong to R&D department, followed by Sales and HR department | Appendix C-3 |
| JobRole | Multiple job roles in different departments. The more populated roles are Sales Executive, Research Scientist and Laboratory Technician | Appendix C-4 |
| Education | Majority possess Education level 3 and 4 | Appendix C-5 |
| Gender | Higher proportion of males compared to females | Appendix C-6 |

### 6.2 Multivariate Analysis

In this section, relationships between multiple variables will be analysed. The multivariate analysis serves as a useful tool to highlight patterns and relationships between different variables. It assists us in filtering out variables that have little to no relationship with the outcome variable and set the stage for the model construction phase. Useful hypotheses can be formed through multivariate analysis coupled with domain knowledge which will help in the feature selection process instead of just lumping all variables into the model.

#### 6.2.1 IBM1: Voluntary Resignation Analysis

It can be observed that even as employees work overtime, their voluntary resignation rate increases **(see Appendix D-1)**. This could possibly suggest that employees become less motivated in completing their tasks at work and also lower satisfaction at work which will eventually lead to higher voluntary resignation.

For the satisfaction surveys, namely, EnvironmentalSatisfaction, JobSatisfaction, JobInvolvement, WorkLifeBalance, we plotted them against Attrition. We noticed a downward trend in all 4 surveys. This means that the higher the score employees give, the lower the number of voluntary resignations.

Based on the JobInvolvement graph, there is a steady increase in the attrition rate as job involvement level decreases **(see Appendix D-2)**. This aligns with our domain knowledge that the extent to which the employee identifies with his / her work is pivotal in determining the employee's likelihood to remain in the job.

Another insight our team has found is that we can clearly see that there is a steady increase in voluntary resignation rate as job satisfaction level decreases **(see Appendix D-3)**. This coincides with our domain knowledge that employees are more likely to voluntarily resign when their job satisfaction level is lower.

With respect to WorkLifeBalance, those with a score of 1 have visibly higher voluntary resignation rates **(see Appendix D-4)**. This is not surprising since work-life balance can ultimately be used as a proxy for job satisfaction. Lower work-life balance serves to show how employees feel about their working conditions and with lower satisfaction levels, there will definitely be a reason to leave the company.

In the case of BusinessTravel, those who travel frequently have a higher voluntary resignation rate **(see Appendix D-5)**. This suggests that some form of relation exists between the frequency in which a person travels and their probability of resignation. A possible reason can be due to the lack of family time during long and frequent overseas trips which is something many do not want to have.

For Employee Source, we noticed that referrals result in a high voluntary resignation rate **(see Appendix D-6)**. However, we would usually expect referrals to be better hires than sourcing from the open market. This suggests that companies would need a more stringent process for evaluating referrals.

For Education level across Departments, both Sales and R&D departments show that the attrition rate decreases when the Education level increases while the HR department showed an opposite trend **(see Appendix D-7)**. This may suggest that some HR employees feel that they are overqualified for their jobs and hence choose to resign.

Generally, when the PercentSalaryHike increases, voluntary resignation decreases **(see Appendix D-8)**. Since the PercentSalaryHike relates to the increase in an employee's income compared to the previous year, a higher PercentSalaryHike would mean that there is a higher increase in the employee's salary. Hence, it would

make sense for the voluntary resignation rate to decrease as the PercentSalaryHike increase.

An interesting insight that our team has found is that voluntary resignation is higher for those with higher monthly income **(see Appendix D-9)**. This is unexpected as using our domain knowledge, income plays a vital role in deciding if an employee leaves or stays in the company. Given that people work to earn income. However, interestingly, the data tells us a different story and with access to more information, we can go a step further and find out why this is so.

Examining MonthlyIncome by Education, our team found out that voluntary resignation is higher for those with higher monthly income having an Education level of 5 **(see Appendix D-10)**. Domain knowledge tells us that individuals with higher qualifications usually earn a higher income. Possible reasons as to why voluntary attrition is higher for this group of people is that they were poached by other companies or their higher qualifications allow them to seek better opportunities in other companies.

### 6.2.2 IBM2: Termination Analysis

From the pool of people who were terminated, we can gather insights into why they were let go.

Based on the graph of PerformanceRating against Attrition, the majority of the employees who were terminated have PerformanceRating of 3 **(see Appendix E-1)**. Even though the PerformanceRating of all the past and current employees are 3 and 4, it can be deduced that these employees were terminated because they are relatively lower performing as compared to the other employees in the company.

Comparing the PriorYearsOfExperience of the current, terminated and voluntarily resigned employees, it can be seen that those who were terminated have relatively lower PriorYearsOfExperience **(see Appendix E-2)**. This could possibly suggest that the employees were terminated because they do not possess sufficient skills, knowledge and expertise needed to perform their jobs.

On closer inspection of the AverageTenure, the team has found that the terminated employees remained with their previous companies for a shorter period of time compared to current employees **(see Appendix E-3)**. This does not directly tell us the reasons why these employees were terminated. A lower AverageTenure could be due to the employee being more frequently terminated or voluntarily resigned from their previous job(s).

## 7 MODEL CONSTRUCTION AND EVALUATION

### 7.1 Logistic Regression Predictive Model: Shortlist Candidates during the Selection Stage

A Logistic Regression predictive model will be constructed to accurately classify current employees and voluntary resigned employees.

In formulating our logistic model, we first used our domain knowledge and narrowed down on possible factors that could possibly affect attrition. This was conducted during the multivariate analysis stage by plotting the factors that we could obtain during the interview process against attrition. Variables that were deemed as having a significant effect on the outcome variable were included in the initial model.

The final model was obtained after various changes were made to the initial model. Each variable was removed one by one and the performance metrics of the model were monitored. This process is iterated until the model with the best performance was achieved. The team used a cut-off of p-value < 0.05 to determine the significance level and ensured that the Odds Ratio Confidence Intervals did not contain 1. Generalized variance-inflation factors (GVIF) of the models were all below 2 which indicated that there were no issues of multicollinearity. Based on the above-mentioned criteria, the final logistic regression model equation is as follows **(see Appendix F-1):**

$$\log_1(\text{Attrition}) = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{DepartmentResearch\&Development} + \beta_3 \text{DepartmentSales} + \beta_4 \text{DistanceFromHome} + \beta_5 \text{JobRoleHuman Resources} + \beta_6 \text{JobRoleLaboratory Technician} + \beta_7 \text{JobRoleManager} + \beta_8 \text{JobRoleManufacturing Director} + \beta_9 \text{JobRoleResearch Director} + \beta_{10} \text{JobRoleResearch Scientist} + \beta_{11} \text{JobRoleSales Executive} + \beta_{12} \text{JobRoleSales Representative} + \beta_{13} \text{MarialStatusMarried} + \beta_{14} \text{MaritalStatusSingle} + \beta_{15} \text{AverageTenure} + \beta_{16} \text{PriorYearsOfExperience}$$

The team had to decide on a threshold level for the logistic regression model. During the model construction stage, the team found out that the accuracy of the model drops when the threshold level is lowered. This was due to the fact that as the threshold level is increased, the stricter the model would be with classifying an employee as likely to voluntary resign. As the threshold level increased, the number of false negatives increased while the number of false positives decreased. The threshold level to be used for the logistic regression model depends largely on our business problems **(see Appendix X)**. Given that we are more concerned about false negatives compared to false positives, the metric that we are more concerned with is recall instead of precision. This is in line with the idea that missing out on potential talented candidates is less detrimental compared to hiring candidates that will eventually resign. Furthermore, the team realised a potential trade-off between precision and recall. A model tuned to attain high precision usually suffers from a

lower recall and vice versa **(Precision and Recall, 2019)**. The team eventually settled on a threshold level of 0.157, which is equal to the proportion of voluntary resignation out of total attritions.

All the models were cross-validated with a train-test split ratio of 70/30. The final model provided a stable accuracy transitioning from the train set to the test set which indicated little to no overfitting issues. After running our logistic model on the test set, we obtained an accuracy of 61.3%, precision of 22.8% and recall of 61.6% **(see Appendix F-4)**. A comparison of the three model performance metrics against the logistic regression model with a threshold level of 0.5 was performed to justify why we selected the model with the lower threshold level **(see Appendix F-5)**.

The team recognised that accuracy is not an appropriate measure for assessing the model performance when the dataset is highly imbalanced. By simply predicting all candidates to stay in the company, very high accuracy can easily be achieved. However, there is little business value in accurately predicting a potential candidate is not going to resign compared to being able to accurately predict a potential candidate is going to resign. This is due to the issue of unequal misclassification costs. Wrongly classifying a candidate that was going to resign is much more costly than wrongly classifying a candidate that was not going to resign.

### 7.2 Classification and Regression Tree (CART): Match Shortlisted Candidates to Relevant Job Openings

The CART model will be used to match candidates to the correct job openings based on the splitting criteria within the decision tree. This would simplify the job of HR staff since there is a suggested list of attributes for them to follow to allocate the new employee to a particular department. Furthermore, given the uniformity of CART, the way the employees are allocated will be unbiased. The decision tree also has high explainability which makes the allocation process simple to execute for new HR staff.

By profiling the current employees who are still with the company against their relevant attributes, we can predict which departments these shortlisted candidates are most suited for based on the information that is available to us in the selection stage. Our team defines most suited by looking at the department in which current employees are in and running the CART model against their attributes.

```
Root node error: 6402/19370 = 0.3305

n= 19370

        CP nsplit rel error xerror     xstd
1 0.247891      0    1.0000 1.0000 0.010226
2 0.007654      1    0.7521 0.7521 0.009396
3 0.005467      6    0.7129 0.7227 0.009270
4 0.001666      8    0.7020 0.7093 0.009210
5 0.000000     11    0.6970 0.7170 0.009244
```

*Figure 1: CP Table for the CART Model*

Due to the large size of current employees, the team decides to set the minimum split to 800 instead of the usual 2. This is to ensure that we get a reasonably sized optimal tree and each terminal node contains a fair percentage of the data. With reference to the plot above, we choose the cp value to be the case when the corresponding test set error is the minimum, which in this case is 0.001666. Hence, we should prune the maximal tree using cp = 0.001666. After pruning the maximal tree, the size of the optimal tree is 8 (Refer to the pdf file attached named "CART Optimal Tree"). After checking the variable importance, we realized that the top 3 most important variables are JobRole, followed by EducationField, and TotalWorkingYears. This suggests that if a shortlisted candidate's attributes cannot reasonably fit into any of the profiles generated by the CART model, the company should pay more attention to these 3 variables when deciding which department to allocate the shortlisted candidates to.

### 7.3 Logistic Regression Explanatory Model: Retain Current Employees

Lastly, the team constructed another Logistic Regression model, plotting Attrition against variables classified as (2) **(see Appendix B)** - variables which the company could reasonably control and influence after the hiring process. The variables identified could be used to pinpoint variables in which employees highly value. This could be used to modify the current company policy in an attempt to retain current employees. Using the insights obtained from exploratory data analysis and our domain knowledge, we created an initial logistic regression model.

The final logistic regression model was obtained after removing variables based on their p-value and Odds Ratio Confidence Intervals that contained 1. GVIF of the models was ensured to be below 2. Based on the above mentioned criteria, the final logistic regression model is (**see Appendix I-1**):

$$\log_2(\text{Attrition}) = \beta_0 + \beta_1 \text{PercentSalaryHike} + \beta_2 \text{BusinessTravelTravel\_Frequently} + \beta_3 \text{BusinessTravelTravel\_Rarely} + \beta_4 \text{OverTimeYes} + \beta_5 \text{EnvironmentSatisfaction2} + \beta_6 \text{EnvironmentSatisfaction3} + \beta_7 \text{EnvironmentSatisfaction4} + \beta_8 \text{JobInvolvement2} + \beta_9 \text{JobInvolvement3} + \beta_{10} \text{JobInvolvement4} + \beta_{11} \text{RelationshipSatisfaction2} +$$

$\beta_{12}$RelationshipSatisfaction3 + $\beta_{13}$RelationshipSatisfaction4 + $\beta_{14}$WorkLifeBalance2 + $\beta_{15}$WorkLifeBalance3 + $\beta_{16}$WorkLifeBalance4

This corroborates with our domain knowledge that the intangible aspects such as work-life balance and job satisfaction are important in increasing employee retention rates. A summary of the logistic model can be found in Appendix I-2.

## 8 PROPOSED BUSINESS RECOMMENDATIONS

The eventual aim is to develop a program that is able to accurately match potential employees with the job openings available by assessing their skillsets, experience and psycho-emotional traits. As a result, the turnover rates can be reduced as much as possible. Our business recommendations will be delivered in the form of a 3-pronged approach, beginning with the hiring process followed by job allocation and finally employee retention. With these recommendations, the team hopes to effectively alleviate the business problems commonly faced by companies in the pharmaceutical industry.

### 8.1 Digitizing the Recruitment Process

From our first analytics solution, the logistic regression model would have given the company knowledge of the variables to look out for in selecting candidates. The team's proposed solution would be implementing a database model to store the resumes provided by all the candidates. This database would act as a digital resume for employers and it would work together with the logistic model so that they can easily sieve out traits, skill sets and experiences determined by the model. Using the results from the database, a ranking system will be implemented to collate the scores of applicants based on how well they 'match' the job opening involved.

Hence, only the top few candidates' resumes would be handpicked to advance to the next stage of the interview process. The shortlisted candidates would already possess the necessary qualifications and ought to be a good match for the job. The company can then move onto the next round of selection, where the company can better understand their psycho-emotional traits and personalities, before making any hiring decisions.

With this solution, the company can determine the best candidates in the shortest amount of time. This would alleviate some burden off the HR department. In addition, this would effectively target the mismatch of skills problem since the logistic model is employed to ensure that the variables present in potential candidates are aligned to what the job role demands. This would effectively lower the cost of the hiring process. Hiring costs are expensive as hiring managers have to interview potential candidates when the time could be spent on doing something more productive and beneficial to the company. Furthermore, money is required for criminal background checks, skills tests, job postings when hiring new people. In the event that a

company hires the wrong person, the same hiring process would have to be repeated, increasing the cost of hiring **(Liz, 2018)**. For example, in the context of the pharmaceutical industry, workers are required to be results-driven. Hence, the logistic regression model will be able to look for like-minded individuals who are suitable for the profile of the job.

### 8.2 Matching Selected Candidates to Suitable Departments using CART

Applying the results from the CART model, the company would know if the applicant and the department/job role applied for is a match. The company could proceed with matches and in the case of a mismatch, propose an alternative to the candidate. So, the next step in the hiring process could be a proposed bootcamp where the candidate could try out employee life in the department for two days or so. This could help in creating a better match since the employee would have the first-hand experience of working in the department. This would further decrease employee attrition due to a mismatch of employee and job opening.

Having the trial bootcamp would also allow the company to get a deeper understanding of the candidate's personality and work ethics on the job. This can further help in matching the right employee for the right job.

At the end of the trial bootcamp, candidates that the company is satisfied with could be offered contracts. This contracts further help the volatility of the departments since employees would have a fixed period of stay with the company. This could be useful in volatile departments like the sales department.

### 8.3 Cultivating Positive Company Culture

Working on the results from the third analytic solution, the logistic regression model would have provided the company with insights on what variables are important to retain employees. The company could then make the necessary adjustments to the organisation culture to ensure a positive environment where employees are comfortable and feel incentivized to stay on. Using our own results as an example, the logistic regression model has singled out relationship and environmental satisfaction, job involvement and work life balance as aspects the company could work on, since the other tangibles, such as salary increases, business travel frequency and overtime work are aspects that a company is unlikely to alter to retain its employee force. It would be more reasonable for the company to act on the following:

For instance, to ensure higher relationship and environmental satisfaction, the company could upgrade its office facilities and add recreational areas for employees to enjoy during break times. The company could also improve by creating a horizontal hierarchy where workers are able to speak their minds. In this case, hot-desking can also aid in enforcing a familial culture. Also, great efforts must be

taken to regularly check-in on the employees and also to realign them back to the company's goals to ensure that they are satisfied with their current job roles and are able to fit into the company culture. Since studies have proven that employees are more productive and make fewer mistakes when they are happy with their working environment, this will only serve to benefit the company in the long run **(Peters, J., 2013)**.

To foster better job involvement, the company could place more responsibility on employees. According to Maslow's hierarchy of needs, self-actualization is the highest level of individuals need. This could be provided to employees with challenging yet rewarding assignments where employees could work at their best and feel a sense of self-fulfillment.

It is imperative that after the company has hired the right person, efforts are taken to retain them. A pharmaceutical company would heavily invest in their employees by sending them to trainings and workshops to increase their potential and productivity in the company. By addressing the variables deemed significant in retaining employees, the company would have improved its business problem of losing investments and human capital since it could have cultivated a group of dedicated and willing employees.

# 9 LIMITATIONS

## 9.1 Limitations of the Dataset

The most obvious limitation is the fact that the dataset is fictitious in nature and our analysis might not generate any meaningful insights in the context of the real world.

Upon looking through the dataset, our team noticed that there is a lack of useful variables as only some of the variables that can be collected during the selection stage are provided to us.

Furthermore, the number of terminations that occurred in the company is very insignificant compared to the number of current employees and those who resigned voluntarily.

Some of the variables are also not defined clearly, as a result, the team would need to rely on our own interpretation of the variables and that may not reflect the actual meaning of the variables.

## 9.2 Limitations of the Models

While the logistic regression model that we chose would have a higher accuracy due to setting a higher threshold level, this is at the expense of overcompensating for the number of employees that will leave the company. This would mean that the

company would have spent less time and resources on training the employees who would stay with the company instead of voluntarily resign.

For CART, the plot of the complexity parameter table constantly suggested a tree with thousands of terminal nodes based on the 1-SE rule. It is not very practical or reasonable to create a decision tree with so many splits most of the time depending on the needs and wants of the company. The fact that we need to adjust the complexity parameter so that we can get a reasonably sized optimal tree suggests that we would need to make our own judgments when deciding on the size of the tree which may be subjective in nature.

For certain categorical variables, the splitting criteria include certain combination of categories that are difficult to interpret or make sense of. For example, one of the splits for deciding attrition status was based on whether the employee rated environmental satisfaction as 2 or 4. The disjoint ratings are not consistent with our domain knowledge. Hence, it is very hard for us to draw conclusive and meaningful insights from these splits.

### 9.3 Limitations of the Business Solutions

First and foremost, the team acknowledges that employee turnover is a natural process and we do not foresee that we are able to eradicate the problem entirely. However, we do expect that our solution and recommendations aid in lowering the employee attrition.

A possible limitation for the first stage of the 3-pronged approach is that it is possible for applicants to lie on their resume. To earn a chance in progressing onto the next stage of the selection process, applicants can include traits, skill sets and experience that they may not have or not very skilled in. This would deprive the other applicants who may be more suited for the job of the opportunity to be interviewed and progressing further into the selection process.

For CART, the model can be quite rigid since it only considers variables extracted from the data set. This would limit the employees' profiles to only a few variables without consideration for other external factors like personality types which are useful for assessing cultural fit. Thus, additional assessment such as a personality test that can provide more valuable information not reflected in the CART model.

Using multiple approaches to retain current employees can incur significant costs. Those people who are tasked with planning and executing these HR analytics projects would have less time to perform more productive tasks that would bring in revenue for the company. Hence, it also incurs opportunity costs. Even if the company executed the suggested approaches to retain current employees, there may be some more significant and deep-rooted issues. One prominent example is office politics. Office politics is a common phenomenon in the corporate world and it

is extremely difficult to solve this issue entirely. So much so that even up till today, large corporations around the world are trying to solve this issue in a variety of ways.

## 10 CONCLUSION

There is no doubt that the mismatch of skills is a huge problem in the pharmaceutical industry and needs to be promptly addressed. Nonetheless, this would not be without its fair share of problems as aforementioned. Hence, with the solutions proposed, the various business problems will be targeted and hopefully, resolved. We do acknowledge that this dataset is a fictional dataset which can result in inconclusive findings at times. Supposed real data is collected, the trends and insights identified may be more pronounced and significant which will greatly help in coming up with better and more specific solutions. Thus, the team believes that the future in Human Resources and ensuring the most efficient hiring process lies in analytics.

## REFERENCES


Aspa, J. (2019, October 23). *Pharmaceutical Industry Overview: Top Regions for Drug Companies: INN.* Retrieved November 1, 2019, from https://investingnews.com/daily/life-science-investing/pharmaceutical-investing/top-pharmaceutical-regions/

Beyond Accuracy: Precision and Recall. (2019). Retrieved November 11, 2019, from https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c

Bouton, K. (2015, July 17). *Recruiting for Cultural Fit.* Retrieved October 7, 2019, from https://hbr.org/2015/07/recruiting-for-cultural-fit

EFPIA. (2018). *2018: The pharmaceutical industry in figures.* Retrieved November 1, 2019, from https://efpia.eu/publications/downloads/efpia/2018-the-pharmaceutical-industry-in-figures/

Hejase, H. J., & Dirani, A. E. (2016, April). *Employee Retention in the Pharmaceutical Companies: Case of Lebanon.* Retrieved November 1, 2019, from https://www.academia.edu/24123031/Employee_Retention_in_the_Pharmaceutical_Companies_Case_of_Lebanon

Hejase, H. J., Dirani, A. E., Hamdar, B, & Hazimeh, B. (2016, April). Retrieved November 2, 2019, from http://www.iosrjournals.org/iosr-jbm/papers/Vol18-issue4/Version-1/H1804015875.pdf

KBManage. (n.d.). *Sales Force Effectiveness.* Retrieved November 1, 2019, from https://www.kbmanage.com/concept/sales-force-effectiveness

Khoele, A., & Daya, P. (2014, August 28). *Investigating the turnover of middle and senior managers in the pharmaceutical industry in South Africa.* Retrieved November 1, 2019, from https://sajhrm.co.za/index.php/sajhrm/article/view/562/774

Liz. (2019, June 27). *9 Benefits to Hiring the Right Candidate the First Time Around.* Retrieved November 1, 2019, from https://www.talentclick.com/resources/9-benefits-hiring-right-candidate-first-time-around/

Megget, K. (2018, January). *Finding pharma's future.* Retrieved November 1, 2019, from http://www.pharmatimes.com/magazine/2018/januaryfebruary_2018/finding_pharm.s_future

Mohan, A. C. (2015, January). *Changing Role of HR Managers in Pharmaceutical Industry.* Retrieved November 5, 2019, from http://globalresearchonline.net/journalcontents/v30-2/13.pdf

Peters, J. (2013, October 8). *The Importance of a Positive Working Environment.* Retrieved November 5, 2019, from https://anz.businesschief.com/leadership/143/The-Importance-of-a-Positive-Working-Environment.

PharamExec. (2010, June). *The New Breed of Leadership.* Retrieved November 1, 2019, from http://files.alfresco.mjh.group/alfresco_images/pharma//2014/08/21/af71497a-d156-439e-8ddf-8304d3bc6770/article-690699.pdf

PWC. (2012, April). *Breaking out of the talent spiral Key human capital trends in Asia-Pacific.* Retrieved November 1, 2019, from https://www.pwc.com/sg/en/breaking-out-of-the-talent-spiral/assets/saratoga_brkgoutoftalentspiral_201204.pdf

Quantzig. (2018, July 30). *Sales Force Effectiveness Pharmaceutical Industry, Pharma Sales Force Effectiveness.* Retrieved November 1, 2019, from https://www.quantzig.com/blog/pharma-boost-sales-force-effectiveness

Reddy, K., & Reddy, K. (2018, May 8). *Why Hiring the Best Candidate is Important? 25 Best Reasons.* Retrieved November 1, 2019, from https://content.wisestep.com/hiring-best-candidate-important/

Severi, T., & Harap, D. (2017). *Pharma's New Hiring Challenges: Positioning Companies for Success.* Retrieved November 1, 2019, from https://www.stantonchase.com/wp-content/uploads/2017/06/SC_WP_HR_Pharma_A4_LR.pdf

Surji, Kemal. (2013). *The Negative Effect and Consequences of Employee Turnover and Retention on the Organization and Its Staff. EJBM. 5. 52-65.* Retrieved November 11, 2019, from https://www.researchgate.net/publication/313636497_The_Negative_Effect_and_Consequences_of_Employee_Turnover_and_Retention_on_the_Organization_and_Its_Staff

The global skills mismatch. (2019). Retrieved November 11, 2019, from https://www.thestar.com.my/news/education/2019/08/18/the-global-skills-mismatch

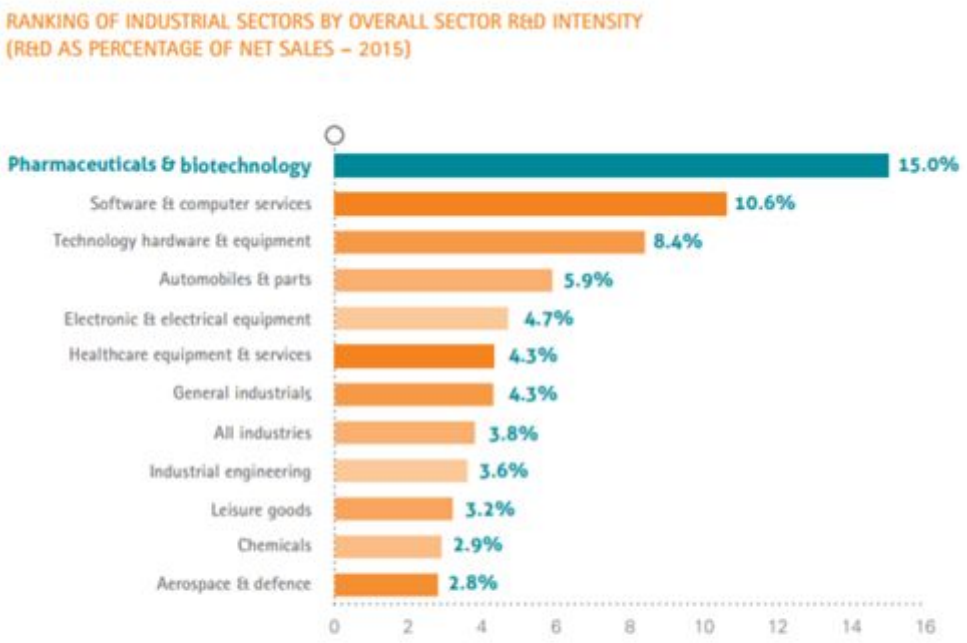
# APPENDICES

*Appendix A-1: Education Requirements for Pharmaceutical Chemists*

| | |
|---|---|
| **Required Education** | Bachelor's degree in chemistry or a related field; research positions often require a master's or doctoral degree |
| **Projected Job Growth* (2018-2028)** | 4% for chemists |
| **Average Salary* (2018)** | $83,850 annually for chemists working in pharmaceutical and medicine manufacturing |

*Source: *U.S. Bureau of Labor Statistics (BLS)*

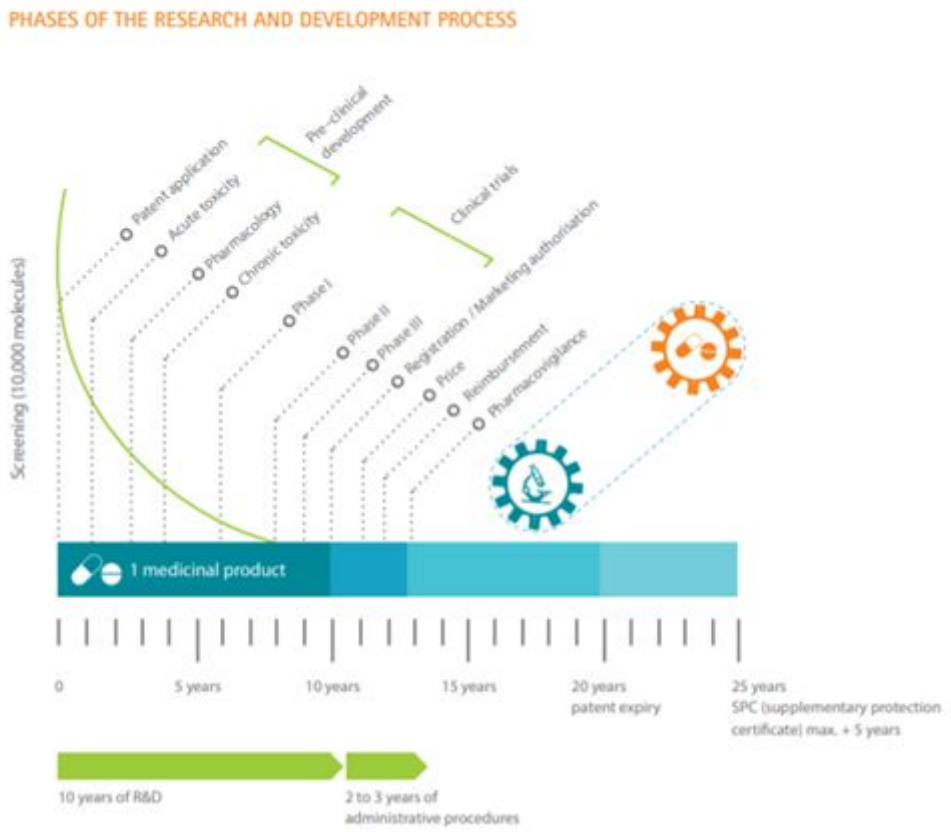
*Appendix A-2: Ranking of Industrial Sectors by Overall Sector R&D Intensity*




*Note:*
*Data relate to the top 2,500 companies with registered offices in the EU (590), Japan (356), the US (837), China (327) and the Rest of the World (390), ranked by total worldwide R&D investment (with investment in R&D above € 21 million).*

*Source: The 2016 EU Industrial R&D Investment Scoreboard, European Commission, JRC/DG RTD*

*Appendix A-3: Phases of the Research and Development Process*



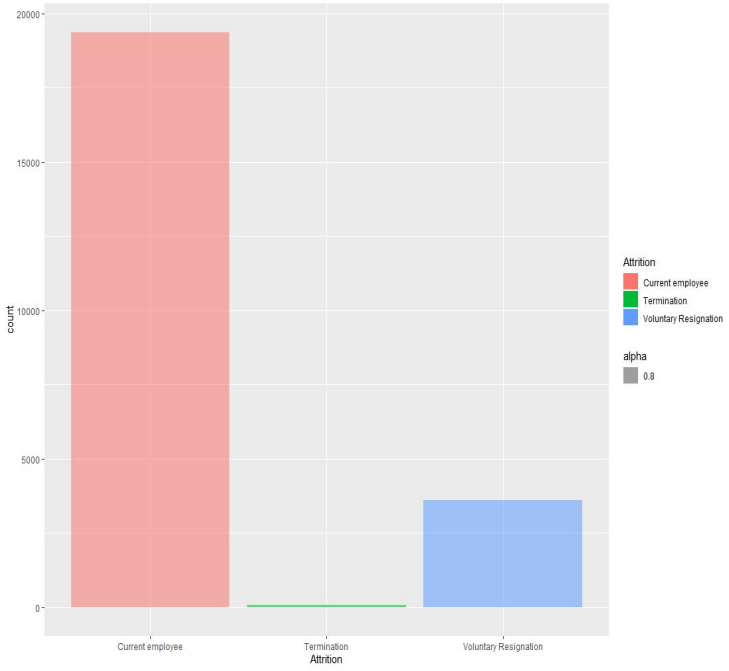

Source: EFPIA member associations (official figures)

*Appendix B: Segmentation of Variables*

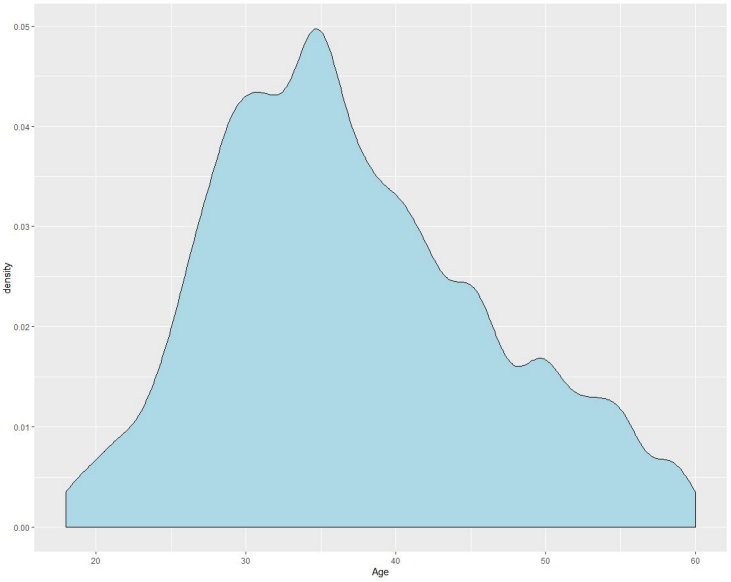
(1) Variables that can be filtered at selection stage
(2) Can be reasonably controlled/influenced by company for current employees

| (1) Variables | (2) Variables |
|---|---|
| Age | EnvironmentSatisfaction |
| Department | JobInvolvement |
| DistanceFromHome | JobSatisfaction |
| Education | RelationshipSatisfaction |
| EducationField | WorkLifeBalance |
| Employee Source | BusinessTravel |
| Gender | OverTime |
| JobRole | StockOptionLevel |
| MaritalStatus | MonthlyIncome |
| AverageTenure | PercentSalaryHike |
| TotalWorkingYears | TrainingTimesLastYear |

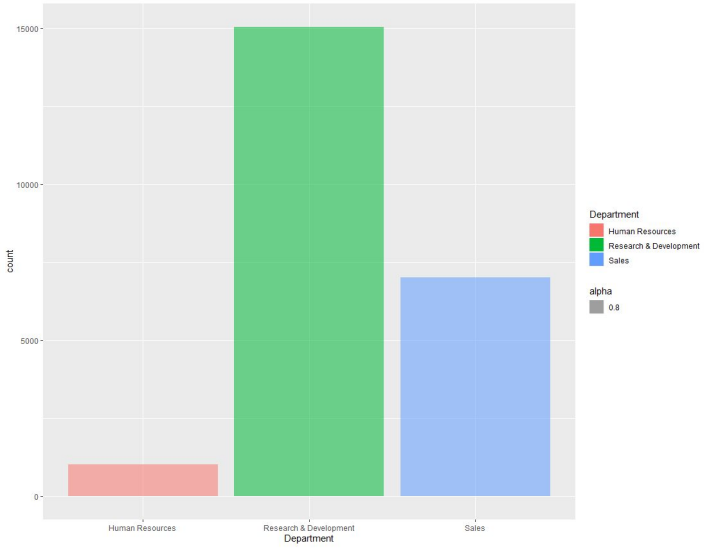
*Appendix C-1: Univariate Analysis of Attrition*



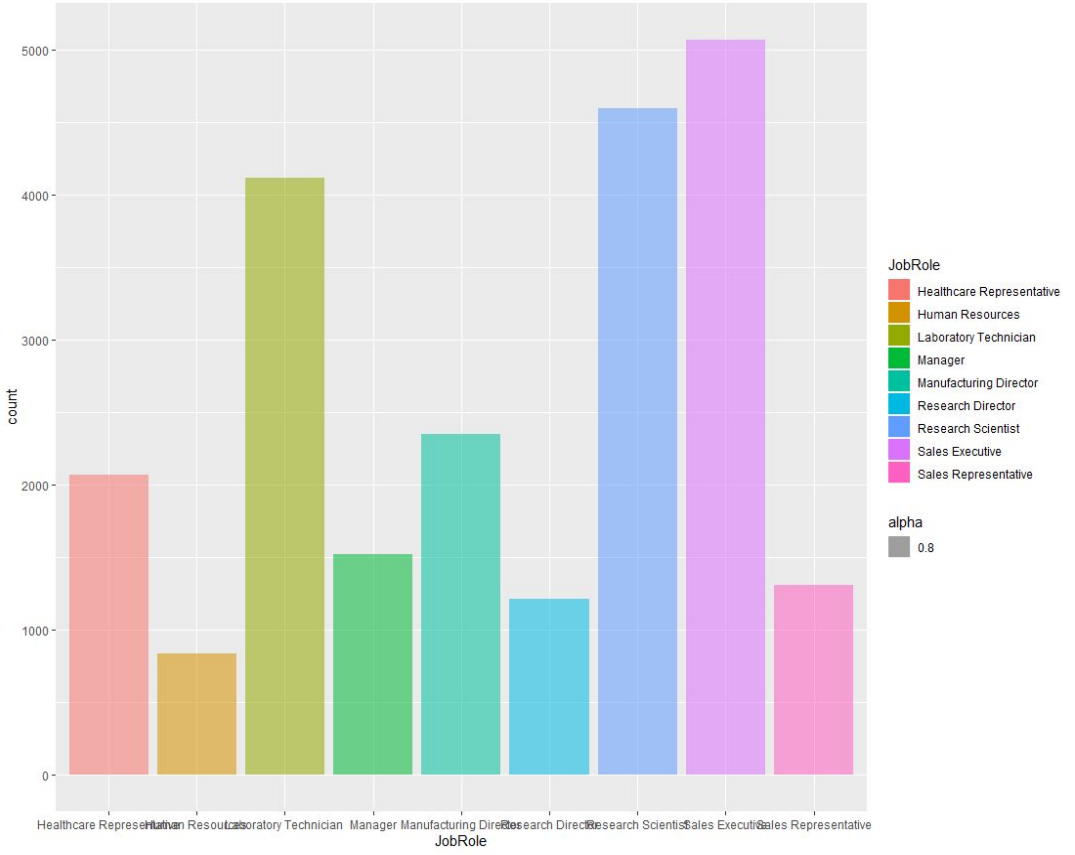

*Appendix C-2: Univariate Analysis of Age*



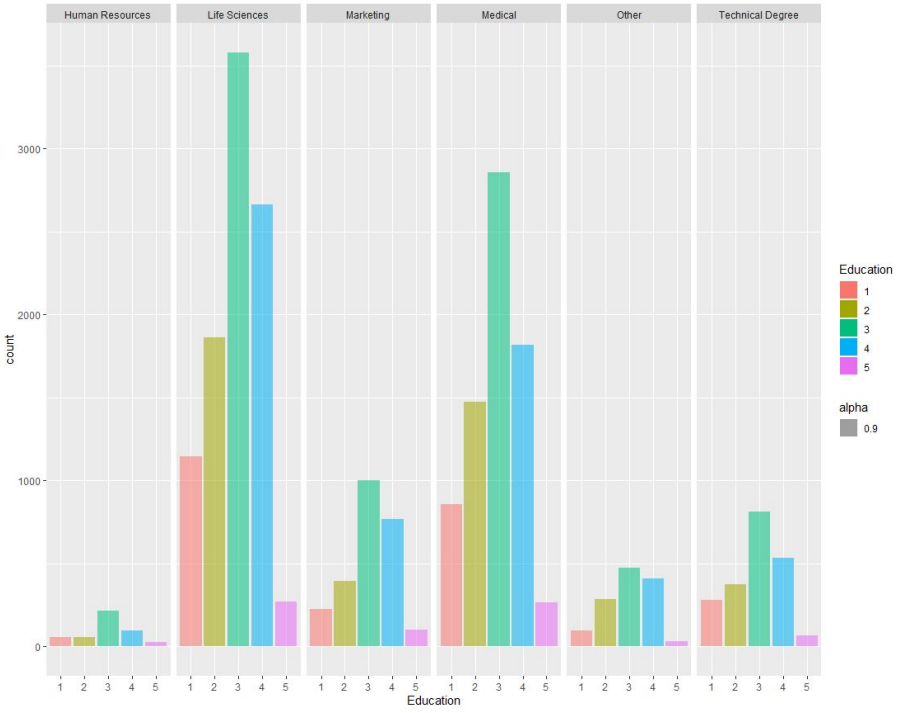

*Appendix C-3: Univariate Analysis of Department*



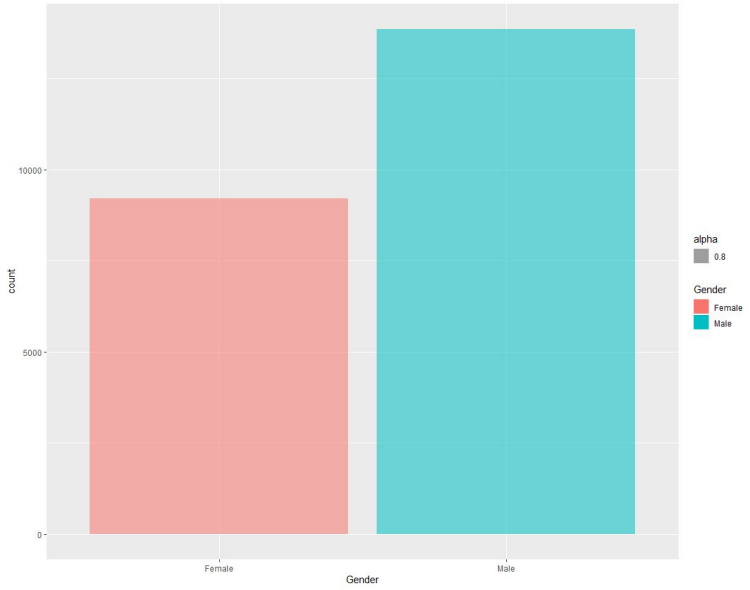
*Appendix C-4: Univariate Analysis of JobRole*



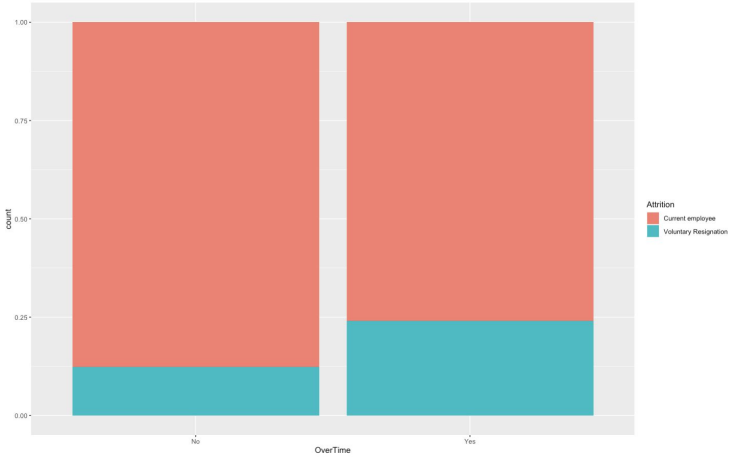
*Appendix C-5: Univariate Analysis of Education, faceted by EducationField*



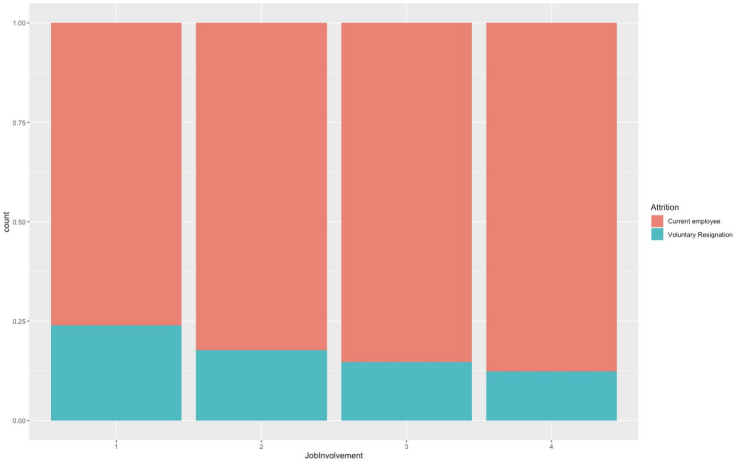

*Appendix C-6: Univariate Analysis of Gender*



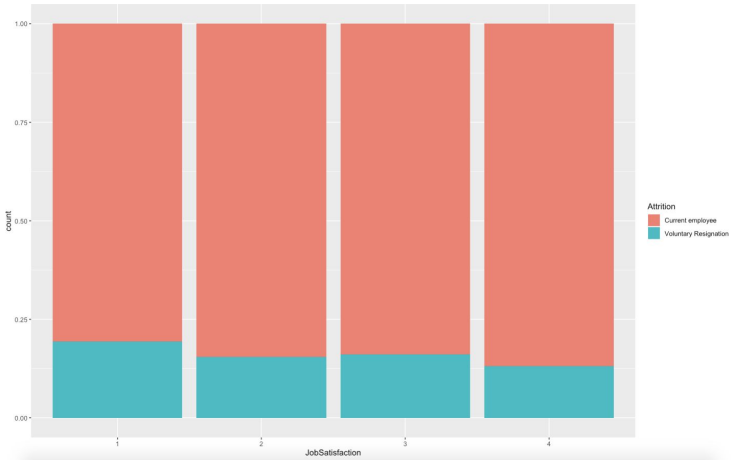
*Appendix D-1: Analysis of Voluntary Resignation due to OverTime*



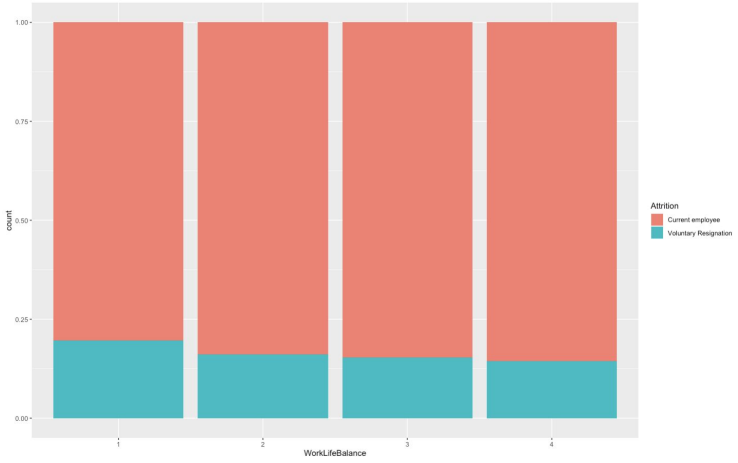

*Appendix D-2: Analysis of Voluntary Resignation due to JobInvolvement*



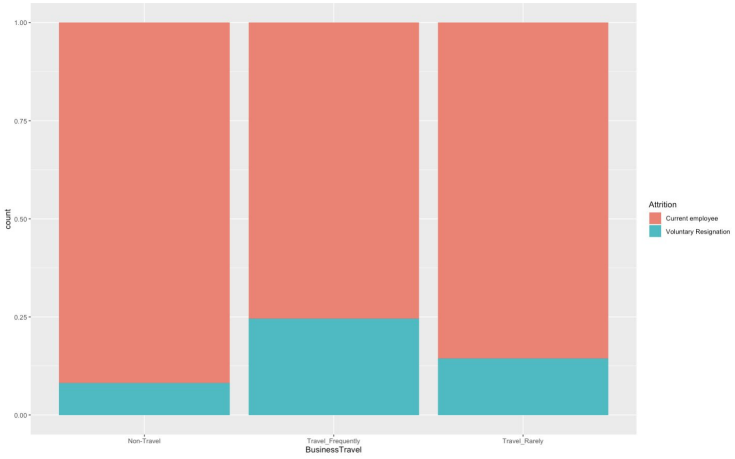

*Appendix D-3: Analysis of Voluntary Resignation due to JobSatisfaction*



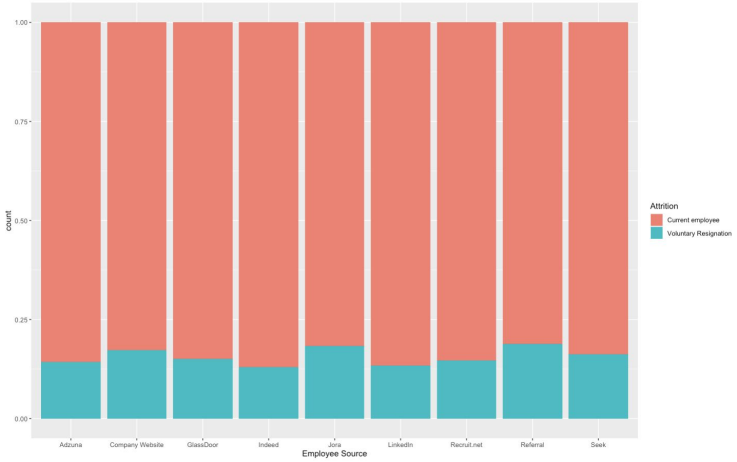

*Appendix D-4: Analysis of Voluntary Resignation due to WorkLifeBalance*



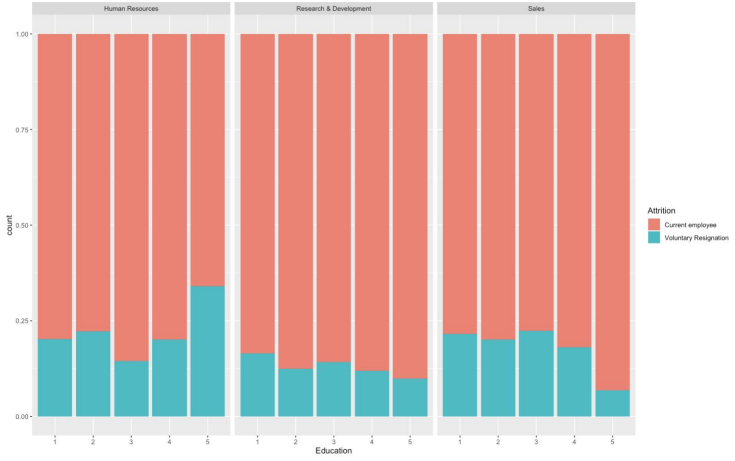
*Appendix D-5: Analysis of Voluntary Resignation due to BusinessTravel*




*Appendix D-6: Analysis of Voluntary Resignation due to EmployeeSource*

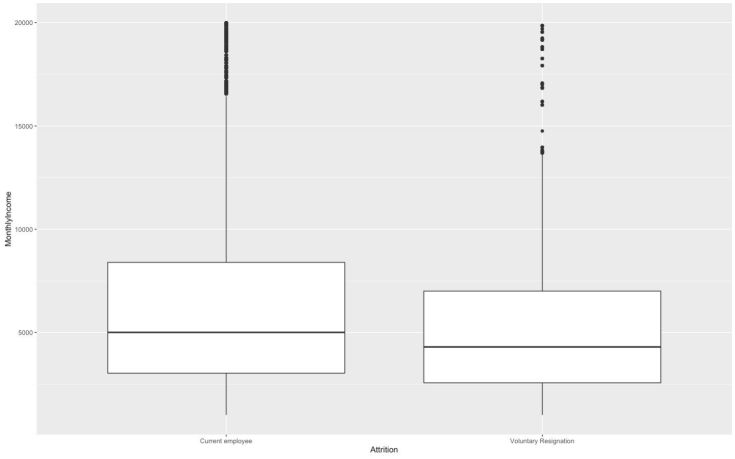



*Appendix D-7: Analysis of Voluntary Resignation due to Education, Faceted by Department*

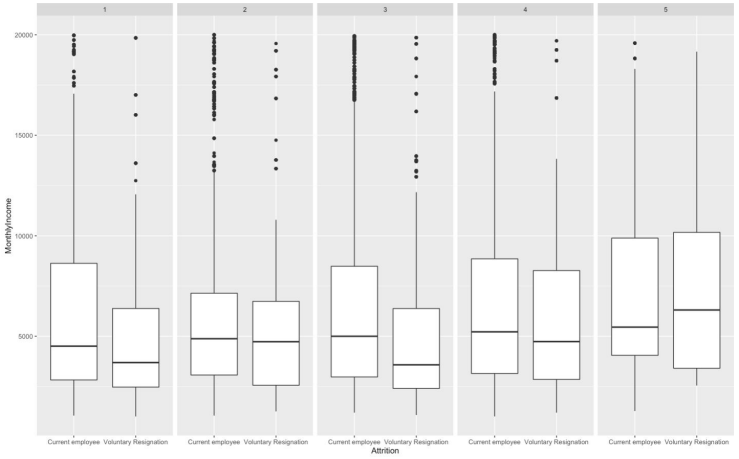



*Appendix D-8: Analysis of Voluntary Resignation due to PercentSalaryHike*

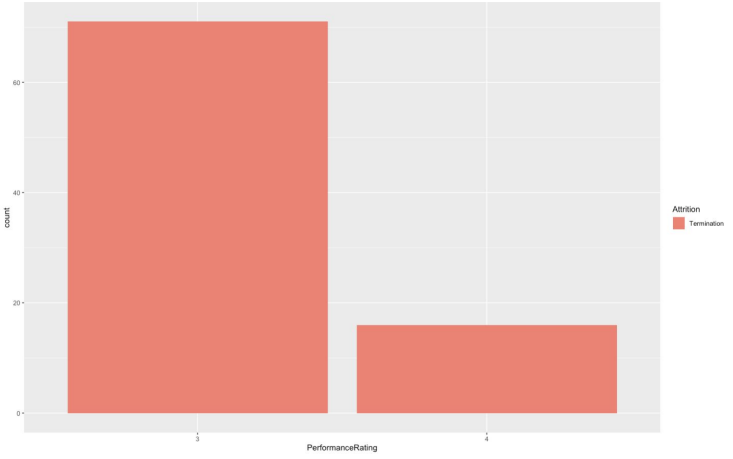
*Appendix D-9: Analysis of Voluntary Resignation due to MonthlyIncome*



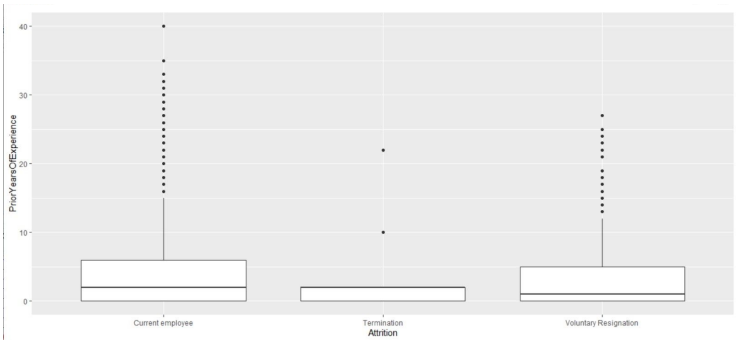

*Appendix D-10: Analysis of Voluntary Resignation due to MonthlyIncome, Faceted by Education*



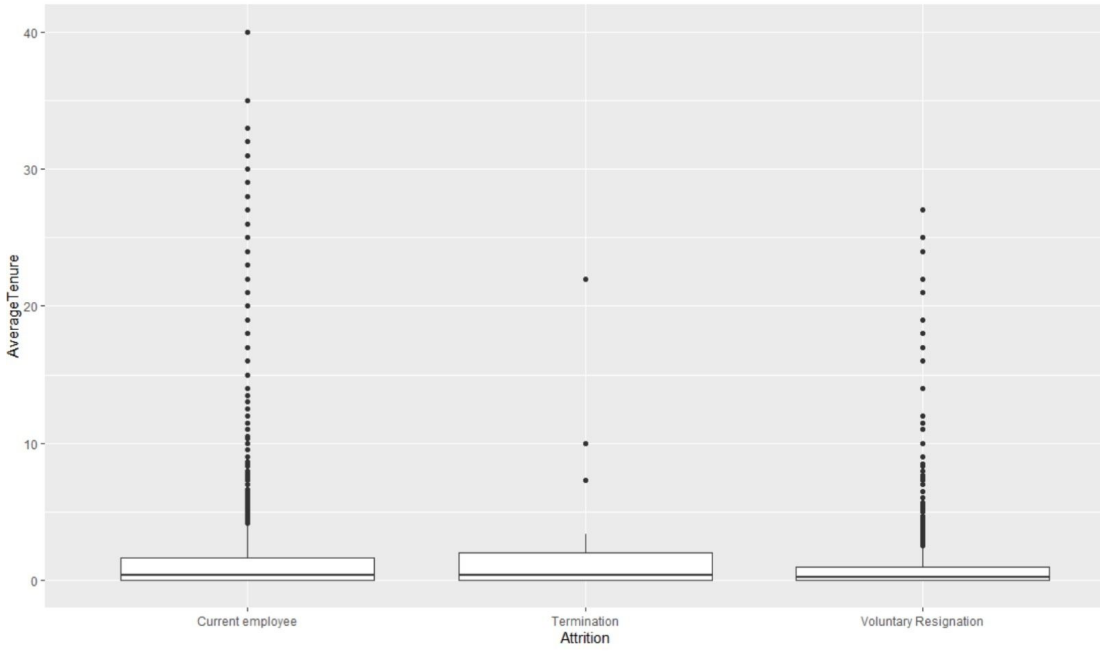

*Appendix E-1: Analysis of Termination due to PerformanceRating*




*Appendix E-2: Analysis of Termination due to PriorYearsOfExperience*



*Appendix E-3: Analysis of Termination due to AverageTenure*



*Appendix F-1: Table of $\beta$ values for Logistic Model 1*

| | |
|---|---|
| $\beta_0$ | -0.42394 |
| $\beta_1$ (Age) | -0.04641 |
| $\beta_2$ (DepartmentResearch&Development) | -0.49514 |
| $\beta_3$ (DepartmentSales) | -0.01194 |
| $\beta_4$ (DistanceFromHome) | 0.02590 |
| $\beta_5$ (JobRoleHuman Resources) | 0.10763 |
| $\beta_6$ (JobRoleLaboratory Technician) | 0.34654 |
| $\beta_7$ (JobRoleManager) | -0.27524 |
| $\beta_8$ (JobRoleManufacturing Director) | -0.11945 |
| $\beta_9$ (JobRoleResearch Director) | -0.40695 |
| $\beta_{10}$ (JobRoleResearch Scientist) | 0.08581 |
| $\beta_{11}$ (JobRoleSales Executive) | -0.00861 |
| $\beta_{12}$ (JobRoleSales Representative) | 0.50062 |
| $\beta_{13}$ (MarialStatusMarried) | 0.14580 |
| $\beta_{14}$ (MaritalStatusSingle) | 0.732012 |
| $\beta_{15}$ (AverageTenure) | -0.03079 |
| $\beta_{16}$ (PriorYearsOfExperience) | 0.02317 |

*Appendix F-2: Summary of Logistic Regression Model 1*

```
Call:
glm(formula = Attrition ~ Age + Department + DistanceFromHome +
    JobRole + MaritalStatus + AverageTenure + PriorYearsOfExperience,
    family = binomial, data = ibm1)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.323  -0.619  -0.497  -0.365   2.706

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -0.51089    0.14893   -3.43  0.00060 ***
Age                          -0.04438    0.00235  -18.85  < 2e-16 ***
DepartmentResearch & Development -0.46510    0.09716   -4.79  1.7e-06 ***
DepartmentSales              -0.00747    0.10034   -0.07  0.94066
DistanceFromHome              0.02414    0.00222   10.89  < 2e-16 ***
JobRoleHuman Resources        0.09974    0.12494    0.80  0.42473
JobRoleLaboratory Technician  0.30710    0.08011    3.83  0.00013 ***
JobRoleManager               -0.42311    0.12347   -3.43  0.00061 ***
JobRoleManufacturing Director -0.07791    0.09472   -0.82  0.41075
JobRoleResearch Director     -0.32751    0.12563   -2.61  0.00913 **
JobRoleResearch Scientist     0.11118    0.07898    1.41  0.15923
JobRoleSales Executive       -0.02453    0.07959   -0.31  0.75788
JobRoleSales Representative   0.48288    0.09556    5.05  4.3e-07 ***
MaritalStatusMarried          0.17125    0.05363    3.19  0.00141 **
MaritalStatusSingle           0.75156    0.05352   14.04  < 2e-16 ***
AverageTenure                -0.01979    0.00943   -2.10  0.03587 *
PriorYearsOfExperience        0.01904    0.00538    3.54  0.00041 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 19951  on 22970  degrees of freedom
Residual deviance: 18703  on 22954  degrees of freedom
AIC: 18737

Number of Fisher Scoring iterations: 5
```

*Appendix F-3: GVIF of Logistic Regression Model 1*

| | GVIF | Df | GVIF^(1/(2*Df)) |
|---|---|---|---|
| Age | 1.11 | 1 | 1.06 |
| Department | 1.72 | 2 | 1.15 |
| DistanceFromHome | 1.01 | 1 | 1.00 |
| JobRole | 2.38 | 8 | 1.06 |
| MaritalStatus | 1.02 | 2 | 1.00 |
| AverageTenure | 2.45 | 1 | 1.56 |
| PriorYearsOfExperience | 2.42 | 1 | 1.56 |

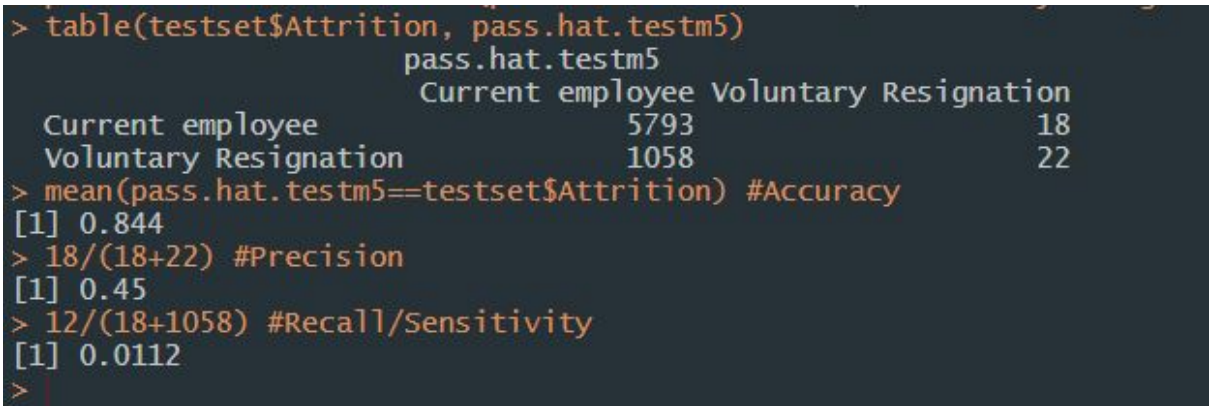
*Appendix F-4: Odds Ratio Confidence Interval of Logistic Regression Model 1*

| | 2.5 % | 97.5 % |
|---|---|---|
| (Intercept) | 0.448 | 0.802 |
| Age | 0.952 | 0.961 |
| DepartmentResearch & Development | 0.520 | 0.761 |
| DepartmentSales | 0.817 | 1.211 |
| DistanceFromHome | 1.020 | 1.029 |
| JobRoleHuman Resources | 0.863 | 1.409 |
| JobRoleLaboratory Technician | 1.163 | 1.592 |
| JobRoleManager | 0.513 | 0.832 |
| JobRoleManufacturing Director | 0.768 | 1.114 |
| JobRoleResearch Director | 0.561 | 0.919 |
| JobRoleResearch Scientist | 0.958 | 1.306 |
| JobRoleSales Executive | 0.836 | 1.142 |
| JobRoleSales Representative | 1.344 | 1.955 |
| MaritalStatusMarried | 1.069 | 1.319 |
| MaritalStatusSingle | 1.910 | 2.356 |
| AverageTenure | 0.962 | 0.998 |
| PriorYearsOfExperience | 1.008 | 1.030 |

*Appendix F-5: Confusion Matrix of Final Logistic Regression Model with Threshold of 0.157*

```
> table(testset$Attrition, pass.hat.testm5)
                     pass.hat.testm5
                      Current employee Voluntary Resignation
  Current employee                3557                  2254
  Voluntary Resignation            415                   665
> mean(pass.hat.testm5==testset$Attrition) #Accuracy
[1] 0.613
> 665/(665+2254) #Precision
[1] 0.228
> 665/(665+415) #Recall/Sensitivity
[1] 0.616
>
```

*Appendix F-6: Confusion Matrix of Final Logistic Regression Model with Threshold of 0.5*



*Appendix G: Linking Classification Model Error Metrics to Business Problems*

| | **Implication** | **Criticality** |
|---|---|---|
| **True Positive** | Correctly predicted employee will leave. | High - Allows company a chance to retain the employees who will leave via early intervention OR avoid hiring such employees in the selection stage to decrease attrition rates. |
| **True Negative** | Correctly predicted employee will not leave. | Low - Generally, not a critical issue for the company. |
| **False Positive** | Predicted employee will leave but did not leave. | Medium - Company might have spent resources trying to retain employee OR did not hire such an employee after this prediction. |
| **False Negative** | Predicted employee will not leave but left. | High - The company would have hired such an employee and exhausted resources on investing in such employees who would leave. This increases attrition rates and hurts the productivity of the company. |

*Appendix H: Variance Importance for the CART Model*

```
> cart3.opt.ibm3$variable.importance
          JobRole    EducationField TotalWorkingYears              Age   Employee Source
      2052.240160        151.044462        134.528068         95.414002         68.973462
    AverageTenure  DistanceFromHome         Education     MaritalStatus
       64.042597          6.672478          3.032957          1.505999
> |
```

*Appendix I-1: Table of $\beta$ Values for Logistic Model 2*

| | |
|---|---|
| $\beta_0$ | -1.04667 |
| $\beta_1$ (PercentSalaryHike) | -0.02139 |
| $\beta_2$ (BusinessTravelTravel_Frequently) | 1.30385 |
| $\beta_3$ (BusinessTravelTravel_Rarely) | 0.63434 |
| $\beta_4$ (OverTimeYes) | 0.82933 |
| $\beta_5$ (EnvironmentalSatisfaction2) | -0.18934 |
| $\beta_6$ (EnvironmentalSatisfaction3) | -0.34926 |
| $\beta_7$ (EnvironmentalSatisfaction4) | -0.35916 |
| $\beta_8$ (JobInvolvement2) | -0.43942 |
| $\beta_9$ (JobInvolvement3) | -0.69452 |
| $\beta_{10}$ (JobInvolvement4) | -0.91573 |
| $\beta_{11}$ (RelationshipSatisfaction2) | -0.36277 |

| | |
|---|---|
| $\beta_{12}$ (RelationshipSatisfaction3) | -0.12133 |
| $\beta_{13}$ (RelationshipSatisfaction4) | -0.21899 |
| $\beta_{14}$ (WorkLifeBalance2) | -0.30392 |
| $\beta_{15}$ (WorkLifeBalance3) | -0.34357 |
| $\beta_{16}$ (WorkLifeBalance4) | -0.35452 |

*Appendix I-2: Summary of Logistic Regression Model 2*

```
Call:
glm(formula = Attrition ~ PercentSalaryHike + BusinessTravel +
    OverTime + EnvironmentSatisfaction + JobInvolvement + RelationshipSatisfaction +
    WorkLifeBalance, family = binomial, data = ibm1)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.293  -0.625  -0.487  -0.399   2.603

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -1.04667    0.15632   -6.70  2.1e-11 ***
PercentSalaryHike            -0.02139    0.00519   -4.13  3.7e-05 ***
BusinessTravelTravel_Frequently  1.30385    0.08475   15.38  < 2e-16 ***
BusinessTravelTravel_Rarely   0.63434    0.07995    7.93  2.1e-15 ***
OverTimeYes                   0.82933    0.03829   21.66  < 2e-16 ***
EnvironmentSatisfaction2     -0.18934    0.05732   -3.30  0.00096 ***
EnvironmentSatisfaction3     -0.34926    0.05248   -6.66  2.8e-11 ***
EnvironmentSatisfaction4     -0.35916    0.05265   -6.82  9.0e-12 ***
JobInvolvement2              -0.43942    0.07642   -5.75  8.9e-09 ***
JobInvolvement3              -0.69452    0.07222   -9.62  < 2e-16 ***
JobInvolvement4              -0.91573    0.09465   -9.67  < 2e-16 ***
RelationshipSatisfaction2    -0.36277    0.06052   -5.99  2.0e-09 ***
RelationshipSatisfaction3    -0.12133    0.05281   -2.30  0.02159 *
RelationshipSatisfaction4    -0.21899    0.05390   -4.06  4.8e-05 ***
WorkLifeBalance2             -0.30392    0.08265   -3.68  0.00024 ***
WorkLifeBalance3             -0.34357    0.07709   -4.46  8.3e-06 ***
WorkLifeBalance4             -0.35452    0.09436   -3.76  0.00017 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 19951  on 22970  degrees of freedom
Residual deviance: 18884  on 22954  degrees of freedom
AIC: 18918

Number of Fisher Scoring iterations: 5
```

*Appendix I-3: GVIF of Logistic Regression Model 2*

| | GVIF | Df | GVIF^(1/(2*Df)) |
|---|---|---|---|
| PercentSalaryHike | 1.01 | 1 | 1.01 |
| BusinessTravel | 1.01 | 2 | 1.00 |
| OverTime | 1.01 | 1 | 1.01 |
| EnvironmentSatisfaction | 1.02 | 3 | 1.00 |
| JobInvolvement | 1.02 | 3 | 1.00 |
| RelationshipSatisfaction | 1.02 | 3 | 1.00 |
| WorkLifeBalance | 1.02 | 3 | 1.00 |

*Appendix I-4: Odds Ratio Confidence Intervals of Logistic Regression Model 2*

| | 2.5 % | 97.5 % |
|---|---|---|
| (Intercept) | 0.258 | 0.476 |
| PercentSalaryHike | 0.969 | 0.989 |
| BusinessTravelTravel_Frequently | 3.127 | 4.360 |
| BusinessTravelTravel_Rarely | 1.617 | 2.212 |
| OverTimeYes | 2.126 | 2.470 |
| EnvironmentSatisfaction2 | 0.739 | 0.926 |
| EnvironmentSatisfaction3 | 0.636 | 0.782 |
| EnvironmentSatisfaction4 | 0.630 | 0.774 |
| JobInvolvement2 | 0.555 | 0.749 |
| JobInvolvement3 | 0.434 | 0.576 |
| JobInvolvement4 | 0.332 | 0.482 |
| RelationshipSatisfaction2 | 0.618 | 0.783 |
| RelationshipSatisfaction3 | 0.799 | 0.983 |
| RelationshipSatisfaction4 | 0.723 | 0.893 |
| WorkLifeBalance2 | 0.628 | 0.869 |
| WorkLifeBalance3 | 0.611 | 0.826 |
| WorkLifeBalance4 | 0.583 | 0.845 |