



AGENO SCHOOL OF BUSINESS

Golden Gate University

MSBA 305-Business Intelligence & Decision Support

Heinz Joerg Schwarz, PhD (abd), MSc.

Spring 2020

Factors influencing Employee Attrition and Retention

Reena Sehitya

SID:0597218



Abstract

The aim of this term paper is to explore factors that affect employee attrition and retention at organizations. I chose this topic for this term project as I always wanted to understand what factors might affect the employee attrition rate at an organization and back that intuition with data analysis. I will analyze factors that affect employee attrition and retention at organizations such as compensation, company culture, learning and development opportunities, and performance and growth opportunities within the company. I will also explore the important question of the various measures a company should take in order to retain employees, what job roles have the maximum/minimum job satisfaction and how the attrition rate is influenced by the employee's personal circumstances so that the suitable personnel are hired in the future. The goal of this paper is to find out the measures that a company can take to avoid attrition of valuable employees, as well as the measures to retain such employees. In order to look for the answer to the above questions, I will perform descriptive data analysis on a fictional data set available from IBM using a BI tool. The study will be conducted through quantitative analysis of the data set and I will also explore what factors influence and predict the chances of turnover for an employee given their details. Using these factors, the goal is to predict the chances of an employee leaving an organization and potentially arriving at actions that the organization can take to retain high performers.

Table of Contents

Introduction	3
Definition of the Employee Attrition dataset	4
Continuous variables.....	4
Categorical variables	6
Problem Statement.....	9
Review of Literature	10
Performance and growth	10
Working Environment	11
Personal factors	11
Method	12
Exploratory data analysis of the Employee Attrition dataset	13
Demographic attribute.....	15
Internal measurement attributes	18
Employee performance attributes	22
Financial related attributes	26
Individual Employee Attribute.....	29
Summary of the Exploratory analysis.....	37
Predicting Employee Attrition at the organization using ML models	39
Step 1: Importing the dataset and relevant libraries in Python	40
Prepare and validate the dataset.....	41
Step 2: Data processing – Encoding category variables and feature selection	43
Step 3: Split the dataset into training and test datasets.....	47
Conclusion	52
Opportunities to improve model performance.....	52
Prescribe a strategy to minimize attrition and retain top performers	54
Key Recommendations.....	55
Conclusion	56
References.....	57

Introduction

With the advent of globalization, people are able to switch their jobs more frequently and easily. Employee attrition rate surges when economy is booming and doing well. However, employee turnover is expensive for an employer. Replacing an employee who resigned costs approximately 21% of their annual salary for an organization (Andrew Chamberlain, 2017). Employee turnover adds to the expenses burden for an organization and also leads to poor employee performance. As per Margaret Rogers (2020), the cost of employee turnover was around \$ 600 billion in the year 2018, and it will increase to around \$680 billion in the year 2020. In their Arts and Social Science Journal article- “Abed Allah, Wasim I Al-Habil and Momhammed Shehadah (2017)” explore broad category of factors that influence employee attrition such as external factors which include perception of the employment, presence of union such as work related factors which also include payments, job performances; and personal factors such as age, gender, distance of the work location from home.

Finally, the focus will shift to exploring effective measures an organization can take to avoid attrition and retain high performing employees. One of the effective solutions mentioned by Margaret Rogers (2020) is to provide effective training and development programs to employees. Other solutions include effective leadership, recruiting suitable employees and improving organizational culture.

Definition of the Employee Attrition dataset

The data set that I will be using for the term project is “IBM HR Analytics Employee Attrition & Performance”. It is a publicly available dataset on Kaggle.com and IBM GitHub. The dataset is a fictional data set created by IBM data scientists. The data set is very clean, and I did not need to do any modification on the data. The data sets are in CSV format with 35 columns and 1470 unique records. Column or attribute is sub-divided into continuous variables and categorical variables. Among 35 attributes, “attrition” is my dependent/target variable and rest of the features/attributes will be considered as independent factors.

Continuous variables

Continuous variables include age, daily rate, distance from home, employee count, employee number, hourly rate, monthly rate, monthly income, number of companies worked, percentage salary hike, standard hour, total years of working, training time since last year, number of years at company, total years in current company, total year since the last promotion, total number of years with current-manager. The definition of all seventeen continuous variables are given below:

- Age: This attribute refers to the age of each employee
- Daily rate: This attribute refers to the income for a contractual hire per day (assumption).
- Distance from home: This attribute refers to the distance of each employee to his/her office from their home
- Employee count: This attribute refers to the count of employee per record
- Employee Number: This attribute refers to the unique identifier for each employee
- Hourly rate: This attribute refers to the income for a contractual hire per hour (assumption).

- Monthly rate: This attribute refers to the income for a contractual hire per month (assumption).
- Monthly Income: This attribute refers to the monthly income of each employee
- Number of companies: This attribute refers to the total number of companies each employee has so far worked
- Percentage salary hike: This attribute refers to the average percentage of salary hike per year of each employee
- Standard hour: This attribute refers to the standard hour each employee works - which is constant value of 80 hours
- Total working years: This attribute signifies the total number of years an employee has worked for.
- Training time last year: This attribute signifies the number of hours an employee spent in training last year
- Years at company: This attribute signifies the average number of years each employee has worked in each company he/she switched
- Years at current company: This attribute signifies total number of years each employee has worked in his/her current company
- Year since last promotion: This attribute signifies total number of years employees have worked since their promotion
- Total Years with their current manager: This attribute signifies total number of years each employee has worked with his/her current manager

Categorical variables

Categorical variables include attrition, performance rating, business travel, department, education field, relationship satisfaction, environment-satisfaction, job level, job role, job satisfaction, job involvement, marital status, over time, education, stock option level, work life balance, over 18, gender. The definition of all eighteen categorical variables are given below:

- Attrition: This attribute with value “Yes” means that the employee has left the company and a value of “No” means that the employee has not left the company
- Business travel: This attribute is divided into sub-categories with “Non-travel” means the employee does not travel for work, “Travel Frequently” means that the employee travels frequently for work, and “Travel Rarely” means that employee travels rarely for work rarely
- Department: This attribute classifies each employee into one of the three departments such as human resources, research and development, and sales
- Education: This attribute specifies the level of education of an employee. The variable can take of the 5 following values - ‘1’ signifies “below college”, ‘2’ signifies “College”, ‘3’ signifies “Bachelor”, ‘4’ signifies “Master” and ‘5’ signifies “Doctor”
- Education field: This attribute classifies each employee into one of the six education fields such as human resources, life science, marketing, medical, other, technical degree
- Environment satisfaction: This attribute specifies the level of satisfaction of employee towards working environment at the organization. The attribute has the following values - ‘1’ signifies “Low” ‘2’ signifies “Medium”, ‘3’ signifies “High”, ‘4’ signifies “Very high”

- Gender: This attribute classifies each employee either as male or female
- Job involvement: This attribute specifies how involved the employee is involved in their job. The attribute can take the following 4 values - '1' signifies "Low" '2' signifies "Medium", '3' signifies "High", '4' signifies "Very high"
- Job level: This represents individual's job level. 1-5, with 5 being the higher job level.
- Job role: This attribute specifies the employee's job roles. The job role is classified into the following 9 categories - Sales Executive, Research Scientist, Healthcare-Representative, Human-Resources, Laboratory-Technician, Manufacturing-Director, Sales Representative, Manager, Research Director
- Job satisfaction: This attribute represents the degree of the employee's job satisfaction. The attribute can take the following values - '1' signifies "Low" '2' signifies "Medium", '3' signifies "High", '4' signifies "Very high"
- Marital status: This attribute classifies each employee into one of the three marital status' such as single, married, divorced
- Over 18: This attribute specifies if the employee's age is above 18. It can take 2 values – "Yes" and "No"
- Over time: This attribute specifies if the employee has done any overtime work. It can take 2 value – "Yes" if the employee has worked overtime and "no" if the employee
- Performance rating: This attribute specifies the average performance rating of the employee. The attribute can take the following values - '1' signifies "Low" '2' signifies "Good", '3' signifies "Excellent", '4' signifies "Outstanding"

- Relationship satisfaction: This attribute specifies the degree to which an individual likes his or her job-related relationships. It can take the following values - '1' signifies "Low" '2' signifies "Medium", '3' signifies "High", '4' signifies "Very high"
- Stock option level: Number indicating what percent of your monthly salary is used to buy stocks at a discount. Values being 0, 1, 2 and 3; with 3 indicating a higher monthly amount.
- Work life balance: This attribute specifies the degree to which the employee is satisfied with their work-life balance. It can take the following values - '1' signifies "Bad" '2' signifies "Good", '3' signifies "Better", '4' signifies "Best"

These are the attribute I will explore to answer how they are leading to employee attrition and later I will explore how each attribute can be used to minimize the attrition rate at the company and retain high performing employees. As of now, I don't see that I need any secondary datasets to supplement the data set that I am going to use in my term project. But if there is a need for supplement datasets, then I will refer to the "Bureau of Economic Analysis" (BEA) website.

Problem Statement

The overall business problem is covered in three sections. In the first section, I am going to explore what factors lead to attrition rate, then I will find out a way to predict attrition rate based on various predictors and then finally I will explore what attributes a company can pay attention to avoid attrition of valuable employees.

At a high level, I am looking to provide answers to the following business problems:

- a. How a HR manager can track attrition within organization by department, various employee personal attributes, job functions, income etc.?
- b. How can a HR manager measure job satisfaction and performance across the organization?
- c. What factors are relevant and related in predicting employee attrition?
- d. What actions and policies a HR manager can take to minimize attrition and retain top performing employees?

Review of Literature

This literature review will discuss three broad categories of factors that influence employee attrition – performance and growth, working environment and personal factors. Within performance and growth category, I will explore factors such as performance review system, learning and development opportunities, career progression, clear job roles and job involvement. For the working environment category, I will discuss factors such as organization culture and structure, relationship of co-workers, and company leadership. As for personal factors, I will explore factors such as age, work-life balance, distance of work from home and remuneration.

Performance and growth

One of the major drivers of employee's quitting their current job at an organization is their lack of recognition in their current role. It is important for an organization to not let workers stagnate in their current roles and promote career growth opportunities for its existing employees so that the employee's odds of leaving the organization does not increase the longer they stagnate in the current roles (Andrew Chamberlain, 2017). According to Dr. Shivani & Dr. Deepa Mishra, top performing employees also place importance on monetary forms of recognition so that they feel that their skills, effort and contributions are valued by the organization. Additionally, employees also consider switching jobs due to perceptions of unfairness with the performance review system (Chowdhury Abdullah Al Mamun, Md. Nazmul Hasan, 2017). Hence, it is critical for an organization to provide clear paths for its employees to grow and regularly progress through the ranks so that they are aware of these career opportunities for growth. Additionally, learning and development opportunities are key for employees to grow within their roles and enable them to progress within the company to higher roles.

Working Environment

Apart from performance and growth, working environment or organization culture also impact crucially in employees' attrition. A healthy working environment influence employee and the way they serve their firm for a longer period of time. The article "Factors affecting employee turnover and sound retention strategies in business organization" mentioned that if the employees are not happy about the culture and working environment then the chances of their quitting their job are quite high (Chowdhury Abdullah Al Mamun Md. Nazmul Hasan, 2017). A firm with healthy working environment attracts talented people to apply for jobs openings at the company. This also improves the performance of existing employees. Culture that promotes innovative and supportive environment also help keeping employees' content towards their organization. Effective leadership is also crucial to increase job satisfaction for employees and employees are inclined to stay if the management and organizational structure is stable which leads to the perception of a friendly working environment. In a friendly working environment, employees see themselves as a contributor towards the achievement of strategic and operating goals of the firm.

Additionally, the perception of co-workers looking for opportunities outside of the company leads to an increase in an employee's turnover and acts as a form of social pressure or rationalization on the employee (Chowdhury Abdullah Al Mamun Md. Nazmul Hasan, 2017).

Personal factors

Personal factors that can affect employees' turnover rate or attrition rate are age, gender, salary, marital status, number of companies worked, distance to the workplace and overall perception of work-life balance. As far as age is concerned, people at the younger age try to frequently change their jobs as compared to the people who are older. In the article "Factors affecting employee turnover and sound retention strategies in business organization", it is

mentioned that employees aged over 30, have certain responsibilities that they consider when looking for another job. In other words, the greater the responsibility on an employee, lesser is the chance of the employee switching jobs. Income is one of the important factors that affects an employees' turnover rate. Higher the income lower is the attrition rate. Andrew Chamberlain in the HBR article, stated that an 10% increment in the base salary can really increase the chances of an employee remaining in the current job role. If any employee gets promoted, the manager must provide a competitive raise in salary in order to retain such a high performing employee. Salary income plays a significant role in retaining talented or result oriented employees.

Method

To answer the proposed research questions, I would utilize the dataset made available by IBM for running descriptive, predictive, and prescriptive analytics. The purpose of this term paper is to explore factors affecting employee attrition. For this study, I chose the data sample from Kaggle.com, and the name of the dataset is "IBM HR Analytics Employee Attrition & Performance". This dataset is a fictional dataset created by IBM data scientists. I will perform descriptive analysis of the dataset using the Tableau BI tool where I will import the data from Kaggle.com. Since, the data is clean, I didn't have to format the data. I will also explore the BI tool to answer important questions such as factors influencing employee attrition, various measures that a company can implement to retain employees, the attrition rate by age, gender and other personal factors. For this study, "IBM HR Analytics Employee Attrition & Performance" dataset is sufficient, and I don't need any supplement dataset/s.

Exploratory data analysis of the Employee Attrition dataset

In this chapter, I am examining how each factor is significantly related to employee attrition. My primary focus here is exploring the data and identifying *trends and relationships of various attributes with attrition*. In this section, I will explore trends and pattern in data through data visualization. In order to accomplish this analysis, I am using Tableau BI tool. The idea here is to create a dashboard that will provide an overview of the attrition rate at the company by employee demographics such as age, sex etc.; financial-related measures such as income, percent hike etc.; performance-related ratings such as job satisfaction, performance rating etc.; employee-category attributes such as education, department etc. and other internal measurements such as years at company, total working years etc.

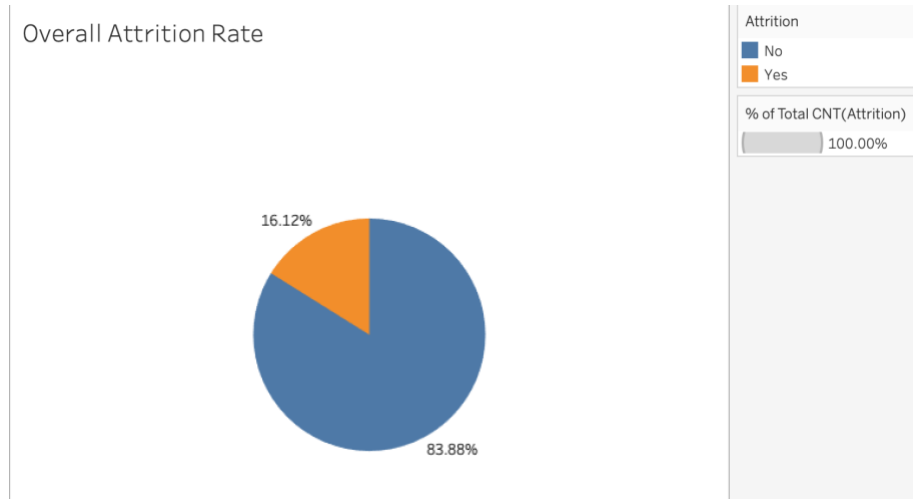
This type of analysis will help me in understanding whether the employees with low job level, low monthly income, low total working years are more likely to leave their jobs or not. It will also help in understanding which department are more susceptible to attrition. Additionally, I will explore what kind of education field such as “Research & Development” or “Technical degree” have higher attrition rate. The overall idea is to frame hypothesis around what factors can be relevant to predict attrition and understand the relationship between the various attributes.

Eventually, all of this will help in understanding what factors are relevant and related to employee attrition which can help in predicting attrition. The factors that are identified to have a strong relationship with attrition will be used as a parameter to run predictive analytics in the next section.

To begin with this chapter, I have uploaded the IBM dataset into the Tableau BI tool and started exploring the data to look for answers of my first business problems through visualizations. My goal here is to help HR managers in tracking attrition within the organization

by demographic attributes, internal measurement attributes, employee performance attributes, financial related attributes, and individual employee attribute.

The overall attrition rate within the organization is **16%** as shown below.



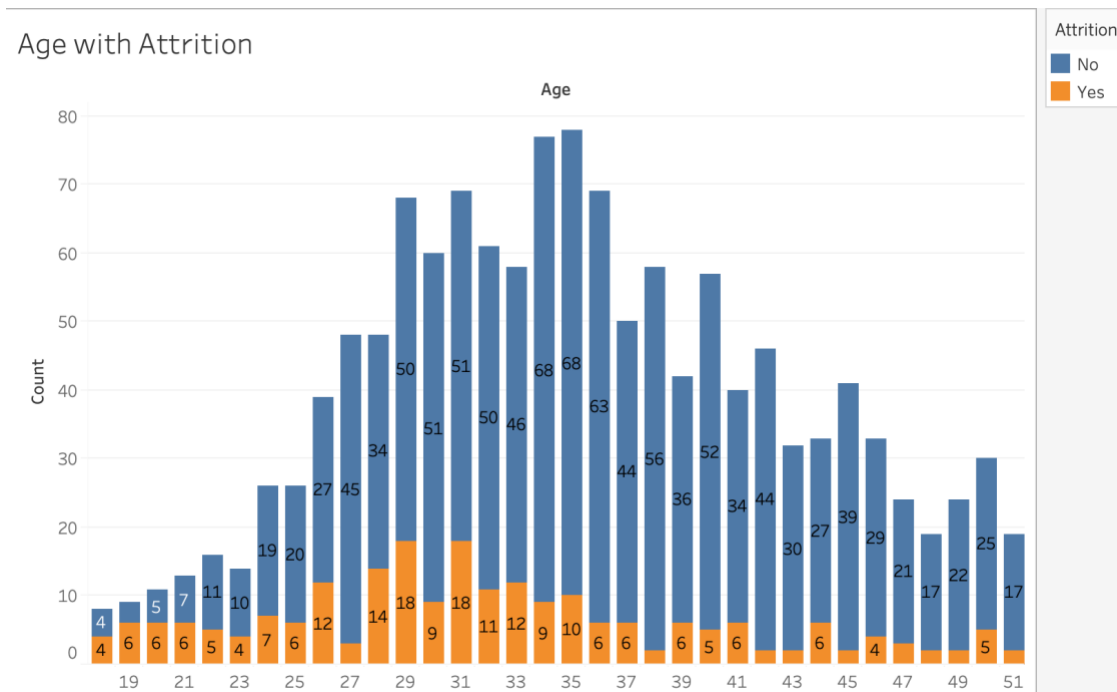
At a high level, all the employee attributes will be categorized into the following 5 themes:

- 1) ***Demographic attributes*** include age, gender, marital status, number of companies worked, employee number, employee count, over 18
- 2) ***Internal measurement attributes*** include total working years, years at company, years in current role, years since last promotion, years with current manager, Training time last year
- 3) ***Employee performance attributes*** include performance rating, job involvement, environment satisfaction, relationship satisfaction
- 4) ***Financial related attributes*** include daily rate, hourly rate, monthly rate, monthly income, percent salary hike, standard hours, stock option level, over time
- 5) ***Individual Employee attribute*** include education, education field, department, job level, business travel, distance from home

I will explore the relationship between attrition and each of the above attribute themes below.

Demographic attribute

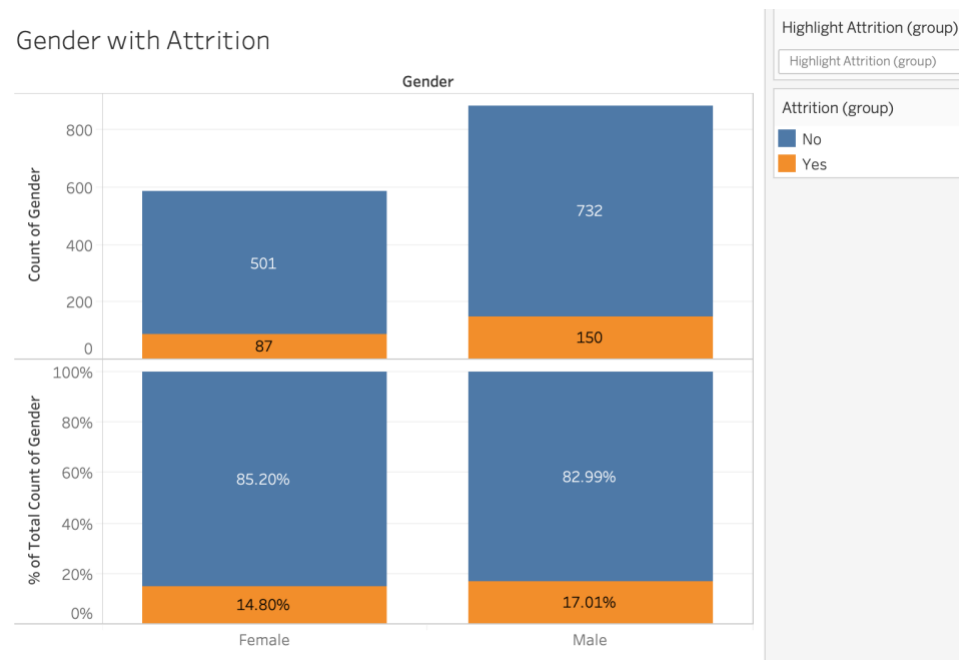
1) Attrition by Age



Analysis Summary: The stack bar above shows that employee attrition is higher for employees who are less than 36 years old, and the attrition rate is low for employees older than 36. The above graph clearly shows that age is negatively related with attrition i.e. as the employee's age increases, the attrition rate decreases. So, the HR managers must specifically focus on employees below the age of 36 below in order to control attrition at the organization.

Key takeaway: Age is negatively related to attrition and younger employees are more likely to quit their jobs.

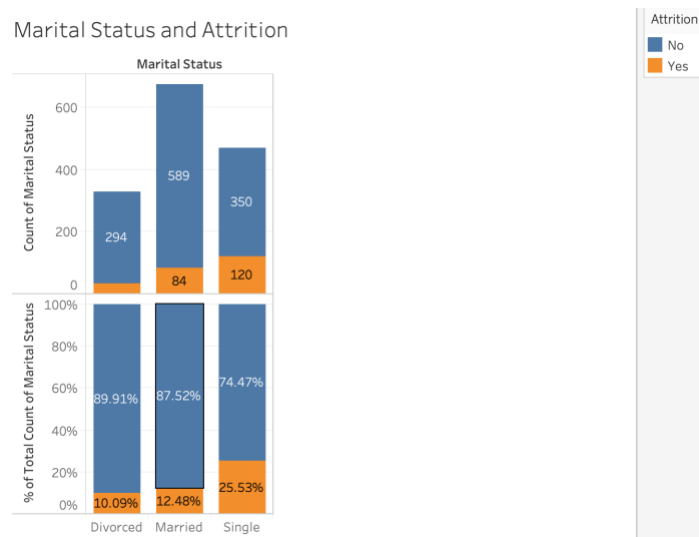
2) Attrition by Gender:



Analysis Summary: By looking at the bar, it is apparent that male has greater attrition rate than female. The attrition rate of female is 14.8% and male is approx.17%

Key takeaway: Male has higher attrition rate than female by almost 2%.

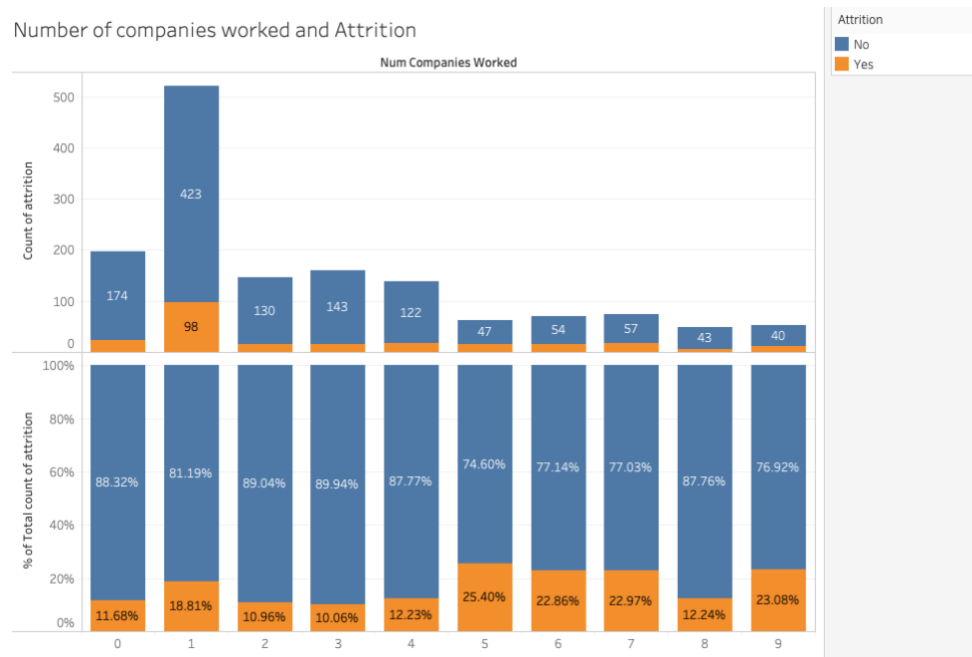
3) Attrition by Marital status:



Analysis Summary: The stack bar above shows that attrition rate is higher for employees who are single followed by married, and then divorced.

Key takeaway: HR managers need to focus on employees who are single since the attrition rate of single employees is almost two times the attrition rate of married and divorced employees

4) Attrition by Number of companies worked:



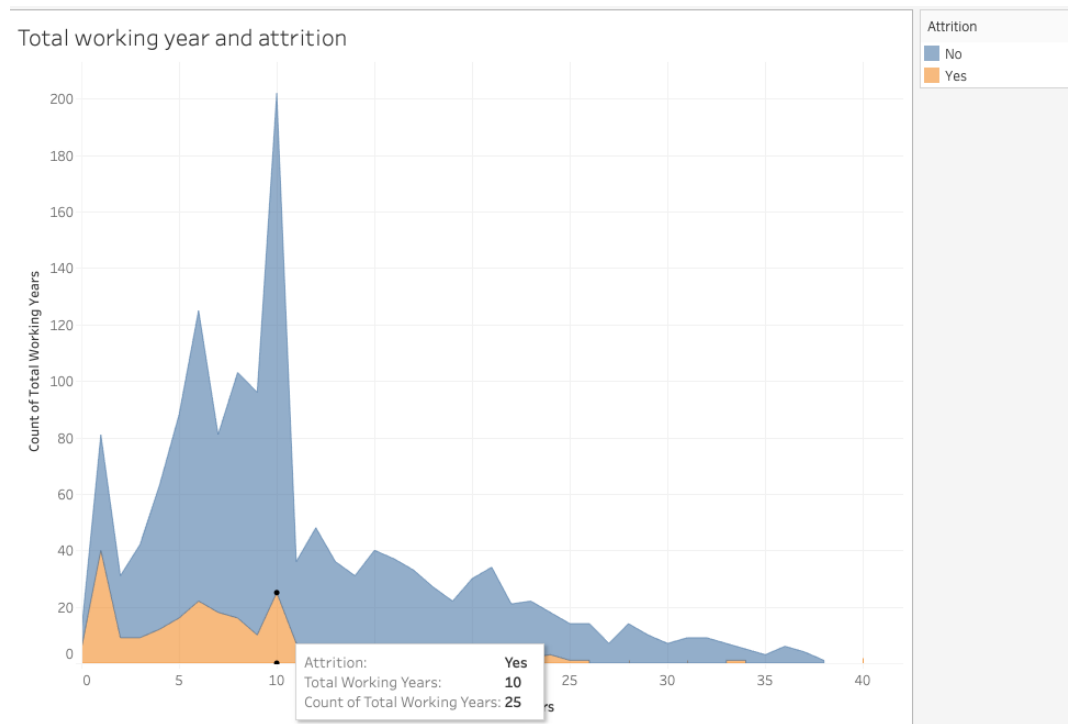
Analysis Summary: The above graph shows that on an average percentage of attrition is greater than 20% as number of companies worked is higher than 4. Whereas, the average attrition rate is around 12% if the number of companies worked is less than 5.

Key takeaway: HR managers need to focus on employees who frequently switched their jobs i.e. more than 5 companies. The overall trend shows that attrition rate is positively related to number of companies worked.

Note: Employee number, Employee count, and Over 18 are the attributes in demographic attributes either have unique values or have same value for every record. So, the analysis of these attributes is irrelevant.

Internal measurement attributes

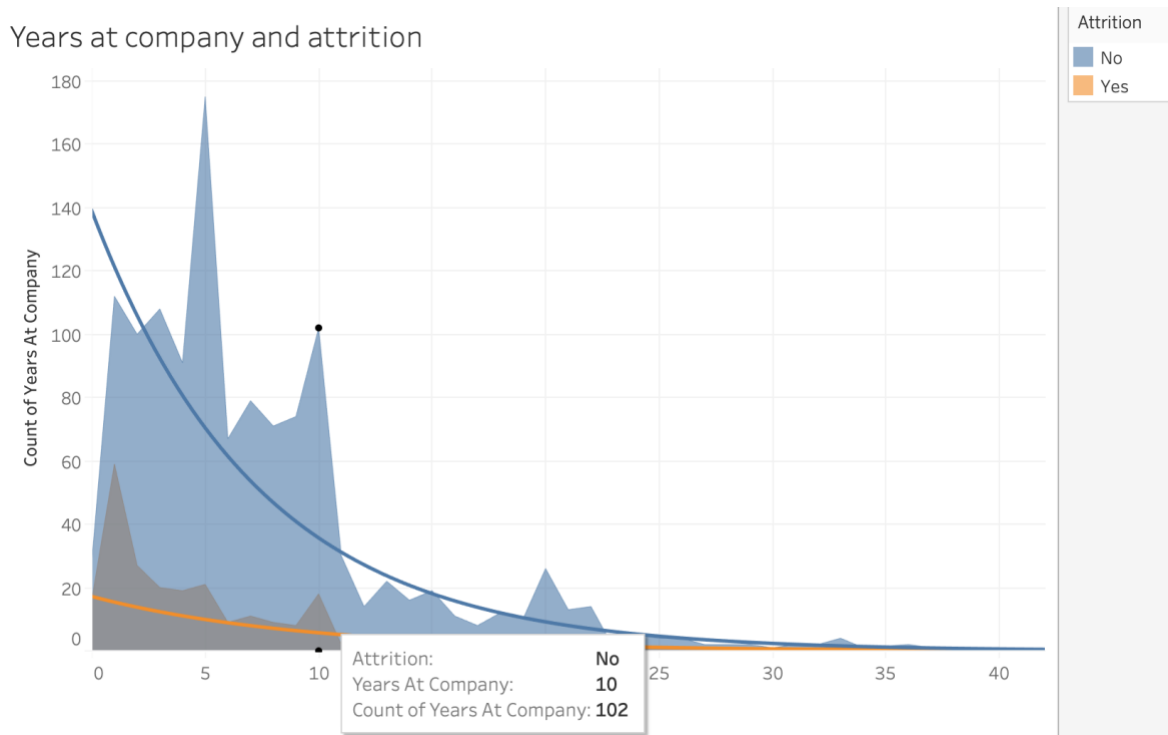
1) *Attrition by Total working years*



Analysis Summary: This area chart above clearly shows that attrition is going down as total working years is increasing. It means that employees who have higher number of working years i.e. > 10 years of working at the company are less likely to quit the company. From 0-10 total working years, overall attrition is spiking at 1, 5, and 10. After 10 total working years, attrition rate is tapering off.

Key takeaway: Attrition rate is inversely related to total working years of employees. HR managers need to focus on employees who have less working experience and especially focus on employees who are nearing their 1, 5- and 10-year anniversaries.

2) Attrition by Years at company

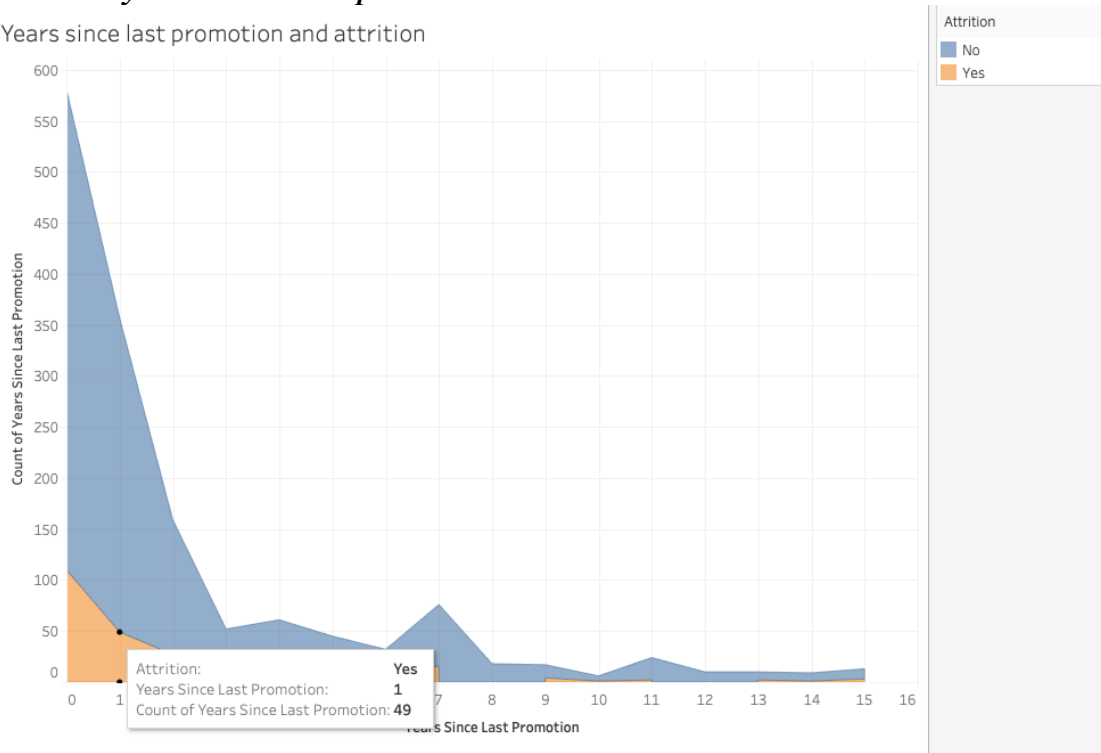


Analysis Summary: This area chart and the trend line above clearly shows that attrition is going down as year at current company is increasing. It means that employees who have worked for more years with the current company are less likely to quit as compared to the employees have worked for less years. Another important attribute that I can see from this area chart is attrition is at its peak for employees who have worked for a year in the current company. Attrition rate is slowing down with employees who have worked for more than 10 years at the current company.

Key takeaway: Attrition rate is inversely related to years at company. HR managers need to focus on employees who have less working experience at the company (<10 years), specially employees who are working for a year for a current company.

3) *Attrition by Years since last promotion*

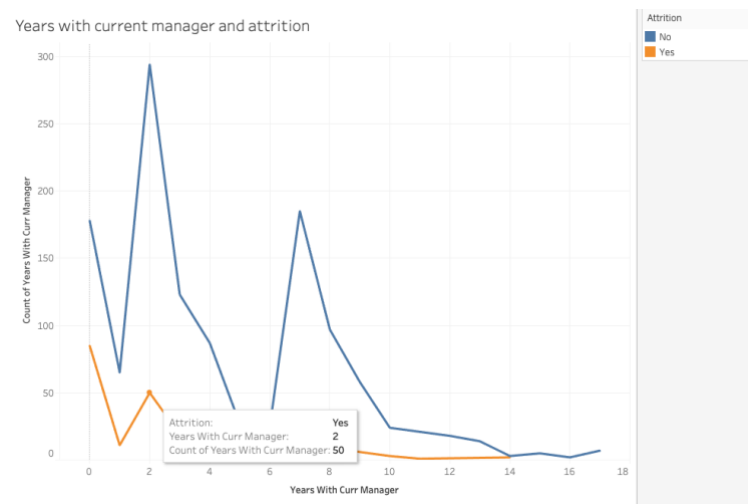
Years since last promotion and attrition



Analysis Summary: This area chart above clearly shows that attrition is going down as year at year since last promotion is increasing. It means that employees who have recently got promoted are more likely to quit the company as compared to the employees who have not got promotion recently. This seems to be strange to observe that employees who have not got promoted for many years are sticking with the current company for a longer period of time.

Key takeaway: Attrition rate is inversely related to years since last promotion. HR managers need to focus on employees who have recently got promoted since these employees are most likely top performers and they are quitting their job at the organization.

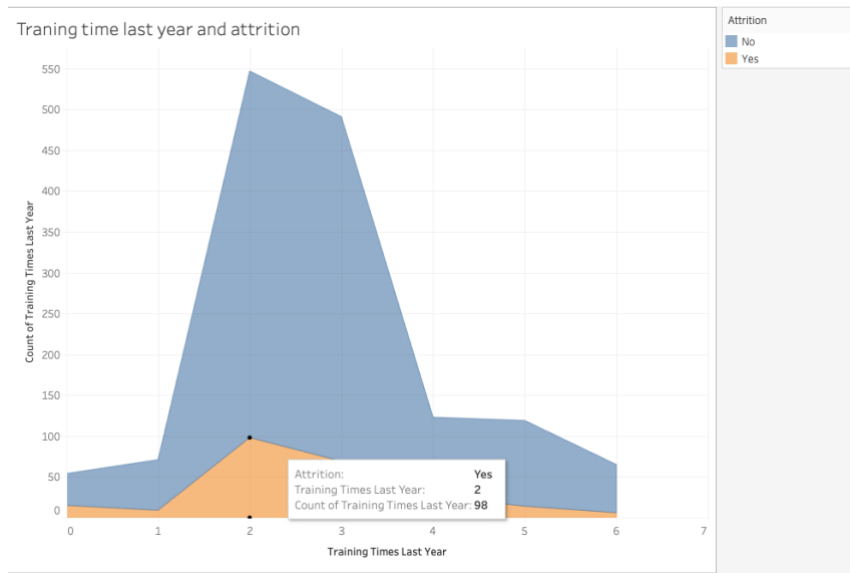
4) Attrition by Years with current manager



Analysis Summary: the line chart above clearly shows that attrition is going down as year(s) at current manager is increasing. It means that employees who have worked for more years with the current manager are less likely to quit the company as compared to the employees have worked for less years. Although, I can see interim spikes of attrition as the years with current manager is going up but overall the trend is downward.

Takeaway: Attrition rate is inversely related to years with current manager of the employee.

5) Attrition by Training time last year

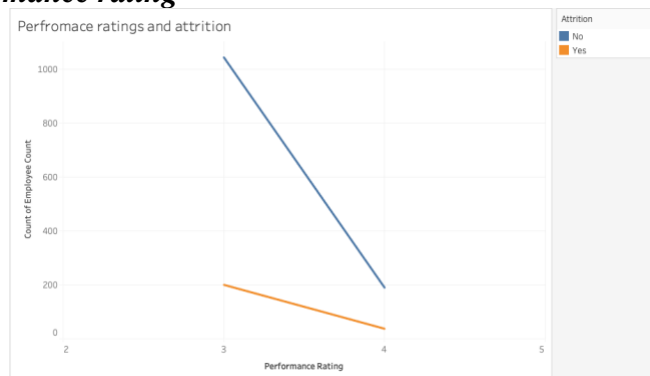


Analysis Summary: This area chart doesn't clearly show any trend of attrition with respect to training time last year. Although, it does show that employees who have training times last year in the range 2 to 4 are more likely to quit the company as compared to employees who have training times last year either less than 2 or greater than 4.

Key takeaway: Attrition is like a bell-shaped curve. There is no clear trend here to focus on.

Employee performance attributes

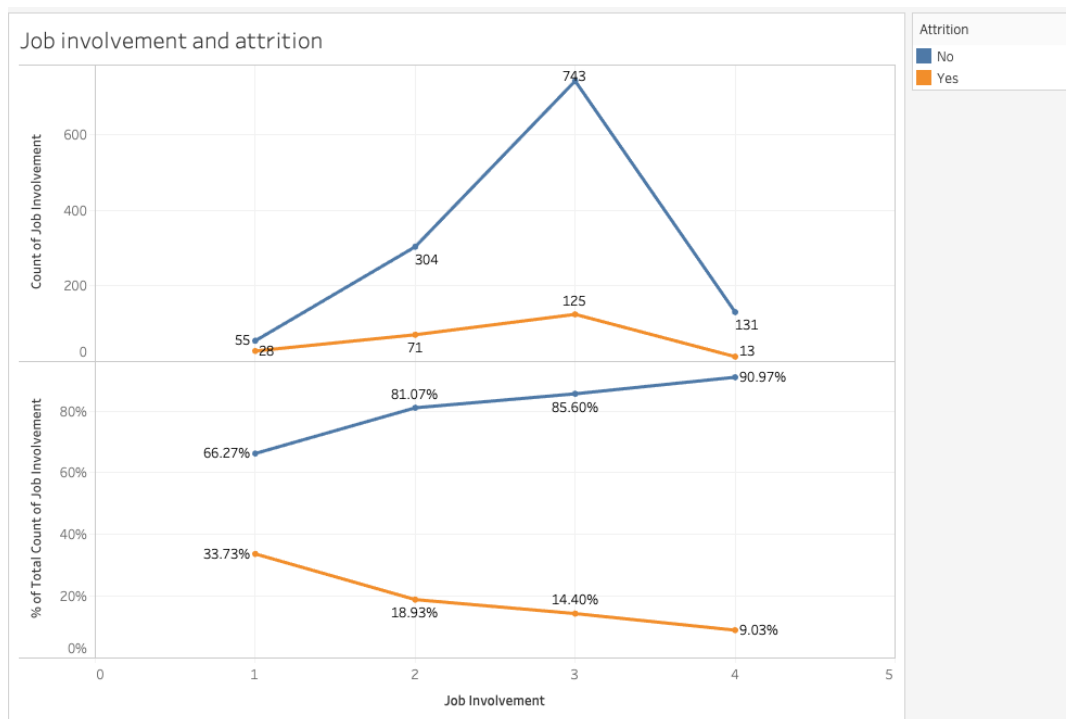
1) Attrition by Performance rating



Analysis Summary: The performance rating data is eye opening even though there are no clear relationships with attrition rate at the organization. The performance rating for all employees at the organization is 3 or 4 and clearly indicates that the performance management system at the company is broken and the system in place is not fair to top performing employees since it does not remunerate them fairly. The only other hypothesis can be that the data is incorrect and cannot be relied upon.

Key takeaway: If the data is accurate then HR should evaluate the efficacy of the performance management system in place at the organization and investigate how top performers feel about the system in place. The performance system looks broken through data analysis and is equipped to reward top performers.

2) Attrition by Job involvement

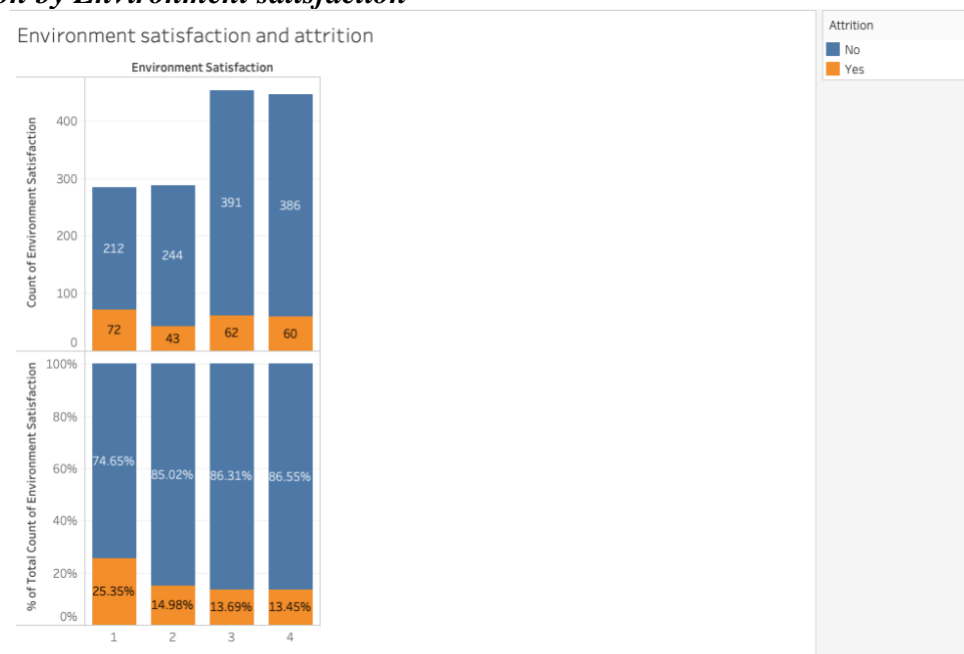


Analysis Summary: This line chart of count of job involvement is showing that attrition is high with job involvement level at 3. The reason that attrition count is high at “job

involvement level at 3” because greater number of employees are falling at this level. If we observe the attrition rate in the graph below, it is showing correct picture that attrition rate is going down from 33.73% to 9.03% with the increase in the job involvement levels from 1 to 4. The trend of attrition rate is going down with the increase in the job involvement levels.

Key takeaway: The attrition rate is going down with the increase in the job involvement levels which is pretty consistent with the hypothesis.

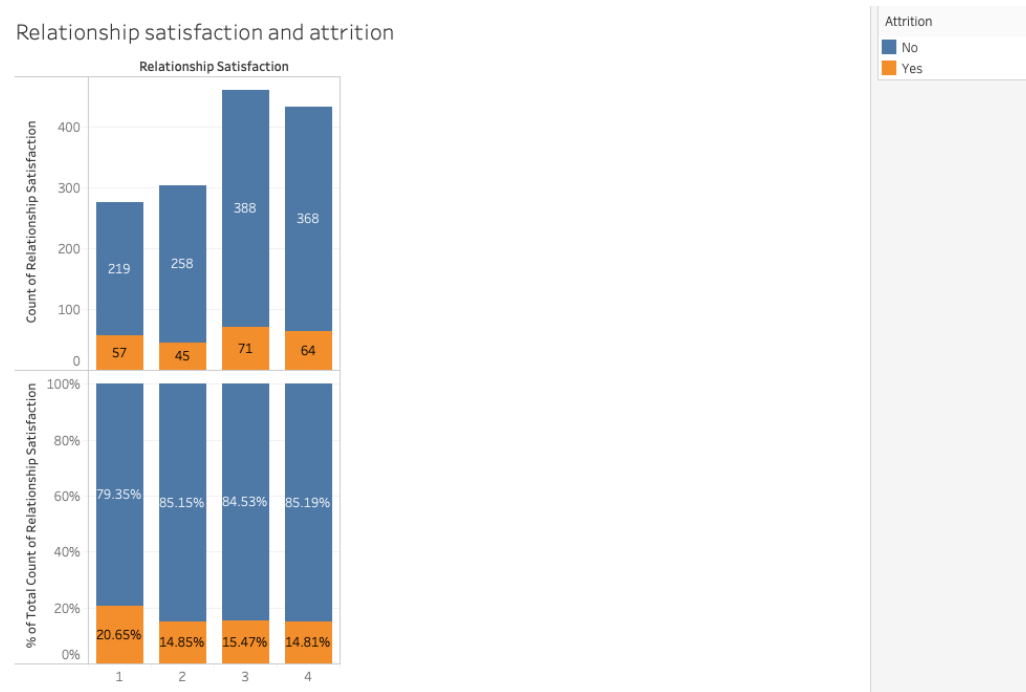
3) *Attrition by Environment satisfaction*



Analysis Summary: This side by side bars is clearly showing that attrition is going down with increasing environment satisfaction level. If we see the percentage wise, it is showing correct picture that attrition rate is going down from 25.35% to 13.45% with the increase in the environment satisfaction levels from 1 to 4. The trend of attrition rate is going down with respect to increase in the environment satisfaction levels.

Key takeaway: HR and people managers need to focus on employees who have low environment satisfaction level. The attrition rate is going down with the increase in the environment satisfaction level.

4) Attrition by Relationship satisfaction

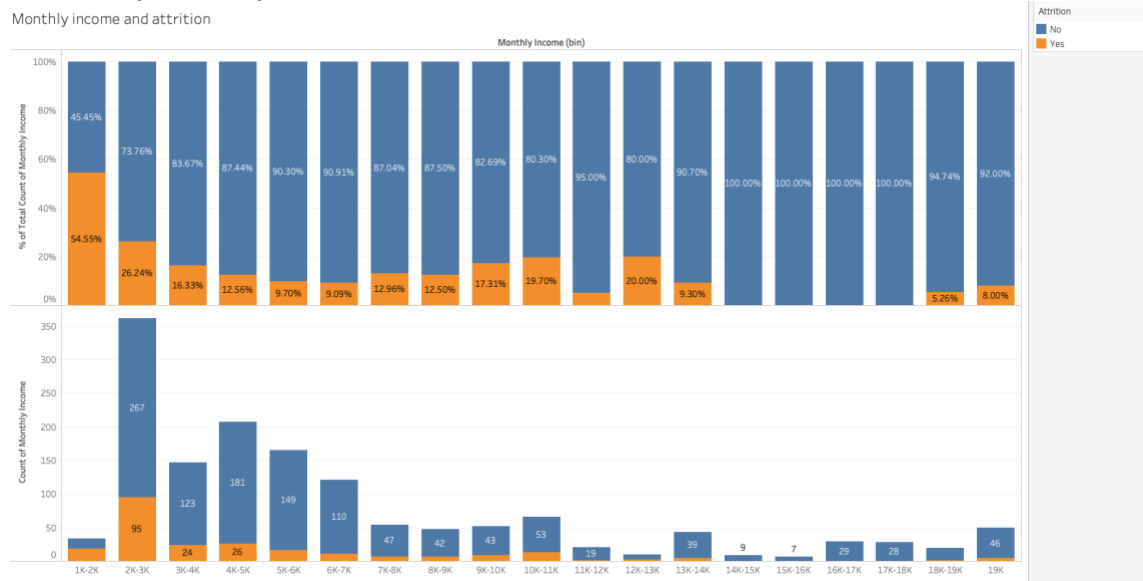


Analysis Summary: This side by side bars is showing that overall attrition is going down with the increase in relationship satisfaction level. If we see the percentage wise, it is showing that the attrition rate is going down from 20.65% to 14.81% with the increase in the environment satisfaction levels from 1 to 4. There is little bit of a spike in the attrition rate at relationship satisfaction level of 3. However, the trend of attrition rate is going down with respect to increase in the relationship satisfaction levels.

Key takeaway: HR and people managers need to focus on employees who have low level relationship satisfaction. The attrition rate is overall going down with the increase in relationship satisfaction level.of a

Financial related attributes

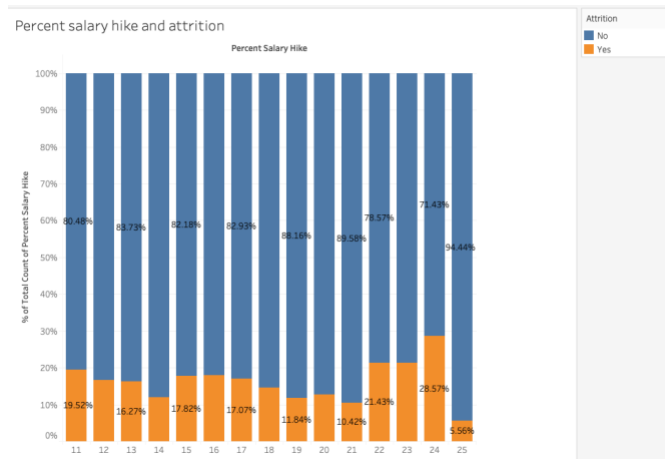
1) Attrition by Monthly income



Analysis Summary: This side by side bars is showing that overall attrition is going down with the increase in monthly income of the employees. If we look at the graphs above, it is fairly obvious that employees who earn < 1K per month have the highest attrition rate as well the number of employees quitting is very high. Additionally, employees < 2k and 3k also have high attrition rates. Above 3k, the attrition stabilizes at around 10% till 13k before finally disappearing completing until employees retire from the company.

Key takeaway: HR and people managers need to assess the remuneration in place for employees earning < 2k and investigate if the pay is below industry standard which might explain the astronomical attrition rate of 29% for such employees.

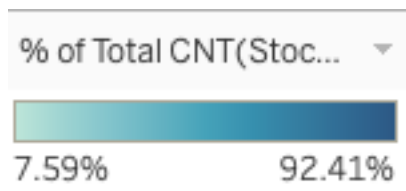
2) Attrition by Percent salary hike



3) Attrition by Stock option level

Stock option level and attrition

Attrition	Stock Option Level			
	0	1	2	3
No	75.59%	90.60%	92.41%	82.35%
Yes	24.41%	9.40%	7.59%	17.65%



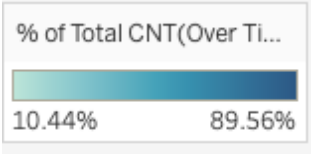
Analysis Summary: This highlight table above is showing that attrition rate is high and in a double digit with the stock option levels of the employees 0 and 3 i.e. 24% and 18%. While the attrition rate is low and in single digit with the stock option levels of the employees 1 and 2 i.e. 9.4% and 8%. This is like a u-shaped curve with higher values at the lowest and highest level of stock option and lower values in the middle level.

Key takeaway: HR managers need to focus on employees who have stock option level of 0 and 4. HR managers need to find out the reason why employees with lowest and highest stock option level are frequently quitting their jobs.

4) *Attrition by Overtime*

Overtime and attrition

Attrition	Over Time	
	No	Yes
No	89.56%	69.47%
Yes	10.44%	30.53%



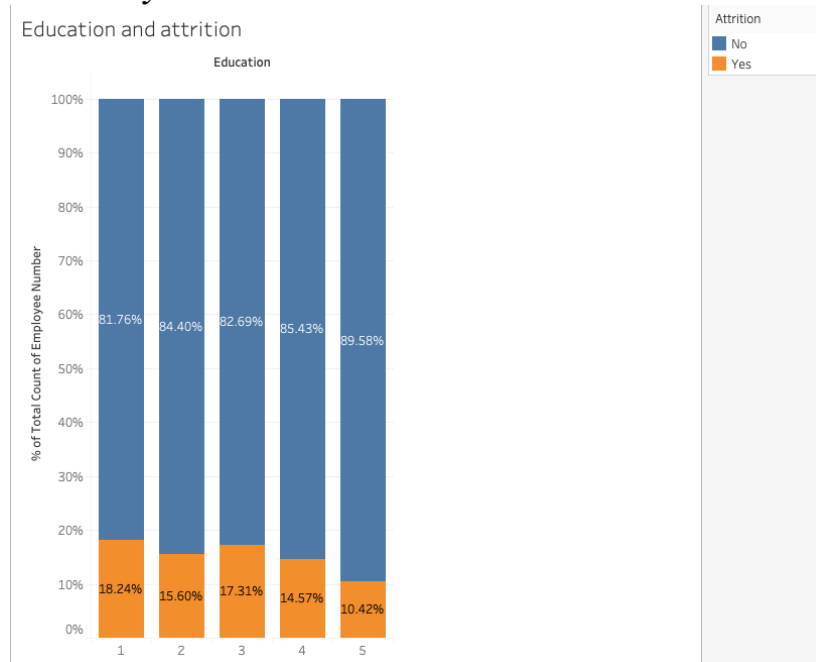
Analysis Summary: The highlighted table above is showing clearly that attrition rate is high for employees who have to do over time. The attrition rate is 31% when employees are working overtime while the attrition rate is comparatively quite low at 10% when employees are not working overtime.

Key takeaway: The attrition rate is higher for employees who work for overtime as compared to employees who do not work overtime. HR managers need to focus on employees who are working overtime.

Note: Attributes such as daily rate, hourly rate, and monthly rate are showing the same trend as that of monthly income after I did the analysis. Since, the standard hour is the same for every record so it was irrelevant for any type of analysis.

Individual Employee Attribute

1) *Attrition by Education*



Analysis Summary: The stack bars above is not showing the trend of attrition rate going down with education level. The attrition rate is quite high at education level 1 with 18.24% followed by level 3 with 17.31%, level 2 with 15.6%, level 4 with 14.57%, and at level 5 with 10.42%.

Key takeaway: There is no trend of attrition rate decreasing with increasing education level. HR managers need to focus on employees with lower education levels (≤ 3) where attrition is higher among employees. HR managers need to investigate why attrition rate is higher at lower education levels.

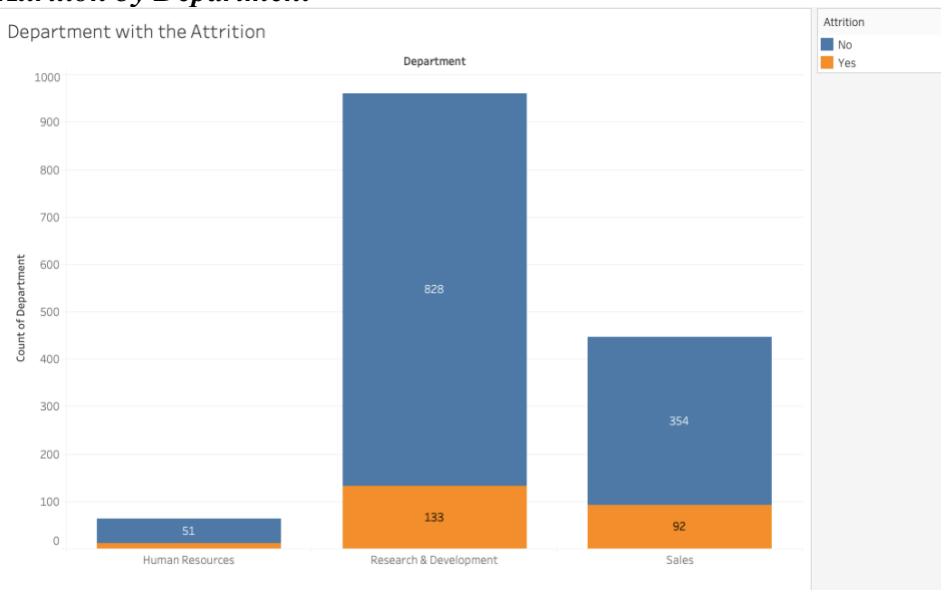
2) Attrition by Education field



Analysis Summary: The stake bars above is showing that the attrition rate is at the highest level with education field of human resources followed by technical degree, life science, medical and then others. The attrition rate is 26% of the employees belong to human resources as an education field then it is 24% of technical degrees and 15% of marketing field. Attrition is lower in employees belong to medical education fields.

Key takeaway: The majority of workers are from the medical, life sciences or marketing field. The only focus area here would be understand the higher attrition rates for marketing employees.

3) Attrition by Department

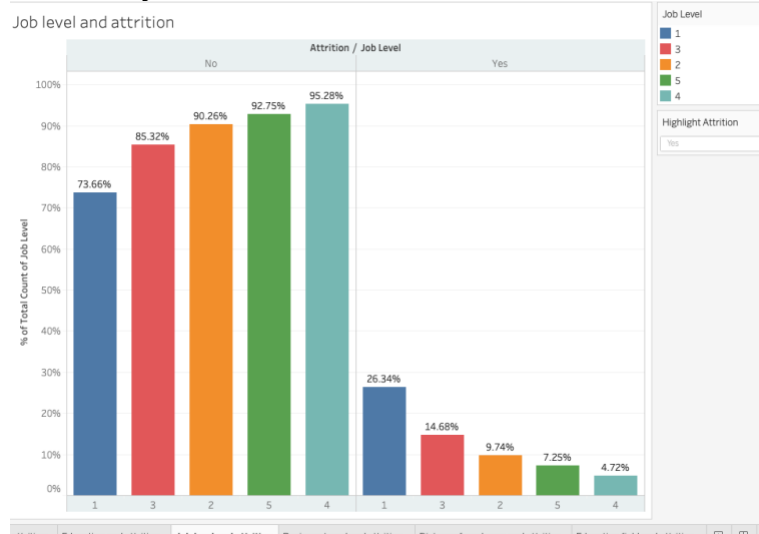


Analysis Summary: By looking at bar graph, it is pretty apparent that Research and Development that has maximum number of employees and has the most attrition followed by Sales and Human Resource. In order to get the correct picture of which department has the most attrition, I calculated the percentage of attrition for each department as shown below:

Department		Attrition Percentage
Human Resource	$12/63 \times 100$	19%
Sales	$92/446 \times 100$	20.60%
Research and Development	$133/961 \times 100$	13.80%

Key takeaway: By looking at percentages, I can conclude that Sales department has the most attrition followed by Human Resource and Research and Development. So, HR managers must focus in the sales department in order to control attrition.

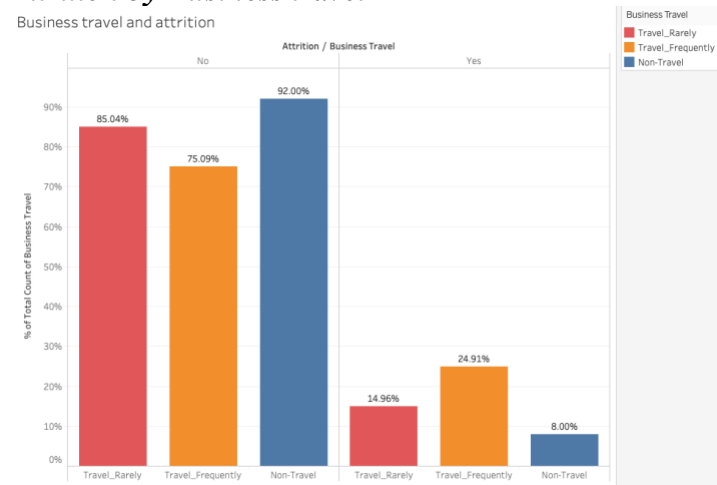
4) Attrition by Job level



Analysis Summary: The side by side bars above is clearly showing that the attrition rate is going down with the increase in the job level. The attrition rate is 26% of the employees with job level 1, 15% with job level 2, 10% with job level 3, 7% with job level 4 and 5% with job level 4. There is definitely downward trend in the attrition level with the increase in the job level.

Key takeaway: Attrition rate is inversely related to job level. HR managers need to focus on employees who are at job level 1 and 2 where the attrition rate is in double digit.

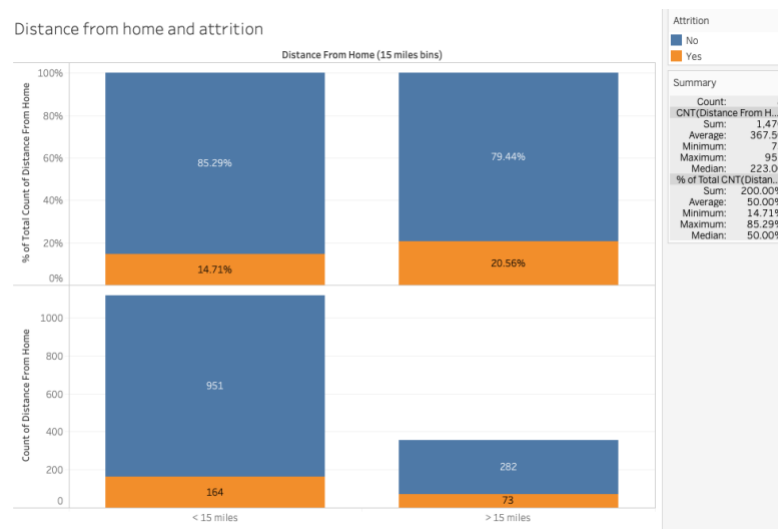
5) Attrition by Business travel



Analysis Summary: The side by side bars above is showing that the attrition rate is at the highest level for employees who travel frequently at 25%. The attrition rate is at 15% for employees who travel rarely followed by 8% for employees who do not have business travels.

Key takeaway: Attrition rate is higher for employees who do business travel frequently. HR managers need to focus on employees who do business travel frequently and understand if it should be an area of focus. Usually, jobs with frequent business travel tend to have higher attrition rates.

6) Attrition by Distance from home

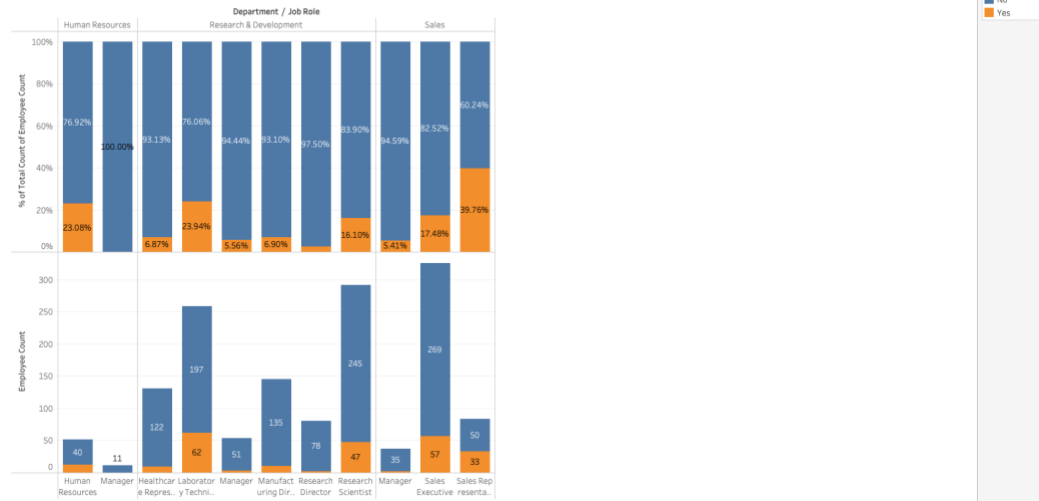


Analysis Summary: For analyzing the impact of “Distance from home” on attrition, I initially binned the “Distance from home” attribute in groups of 3, 10 and 15 miles. However, later on the most appropriate way to bin the “Distance from home” attribute was to bin it in groups of 15 miles. From the stack chart about, it is clear that “Distance from home” does play an impact on attrition and employees staying more than 15 miles away from office have an attrition rate that is ~6% higher than those who live < 15 miles away from office. However, only 24% of the employees stay more than 15 miles away from office.

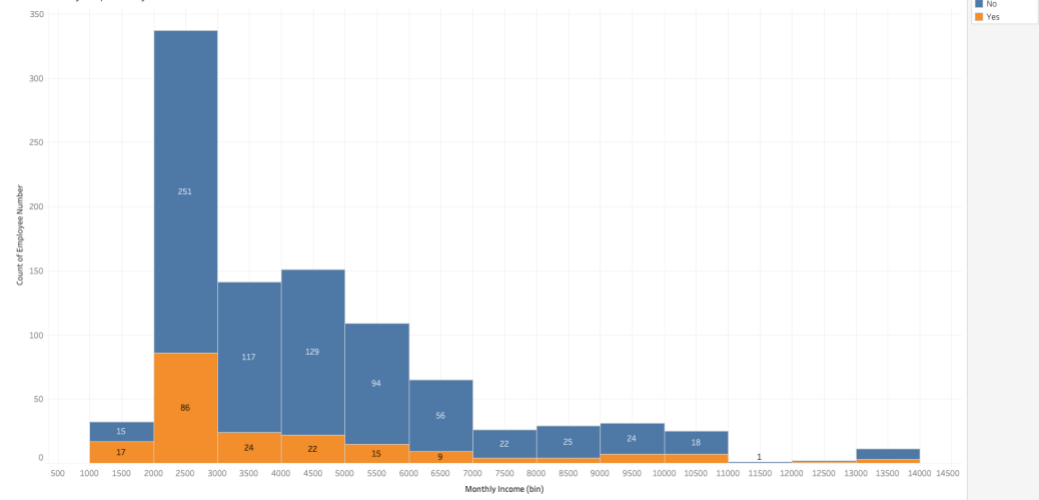
Key takeaway: As the distance from office is > 15 miles, an employee is more likely to quit their job.

Attrition by Department & Job role

Attrition by dept and job role



Attrition by dept and job role

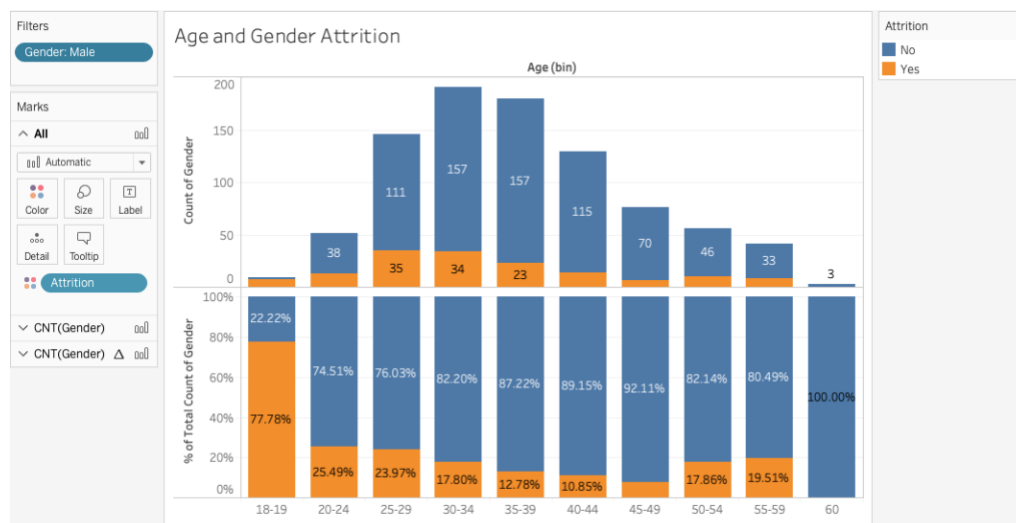


Analysis Summary: After analyzing the impact of “Department” on attrition, I have found that Sales department has the most attrition rate followed by Human Resource and Research & Development. I dig deeper to find out which job role under each department has the most attrition rate. After analyzing job roles in each department with attrition, I found that “Sales Representative” and “Sales Executive” in the sales department has the highest attrition rate of around 40% followed by 24% for “Laboratory technician” in the human

resource department. Job role of “Manager” in every department has the least attrition rate. On further analysis for the salaries of these roles, it is clear that the attrition is primarily driven by lower salaries for these roles.

Key takeaway: It is clear that most employees in the organization are “Sales executives”, “Sales representative” or “Laboratory technician” and that they have remuneration on the lower end. It is important to investigate if salaries in these roles are below the industry standard which might be driving attrition higher for these roles.

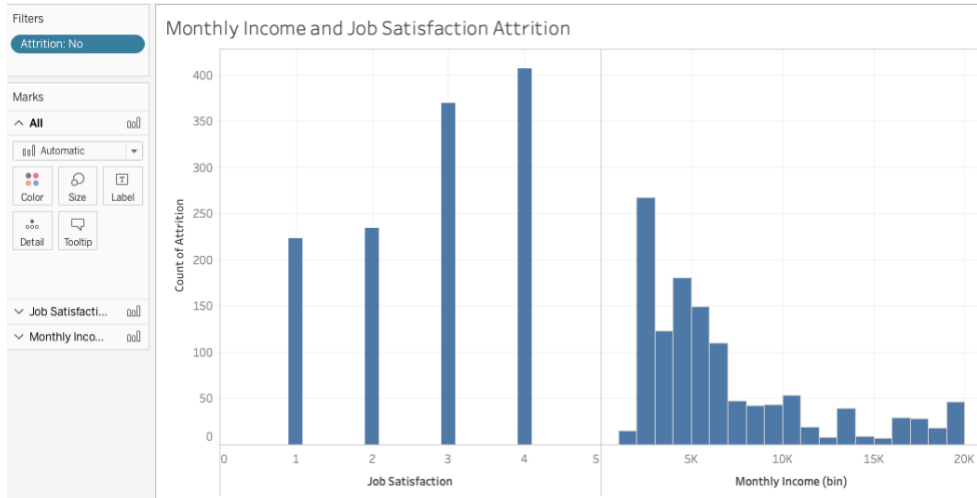
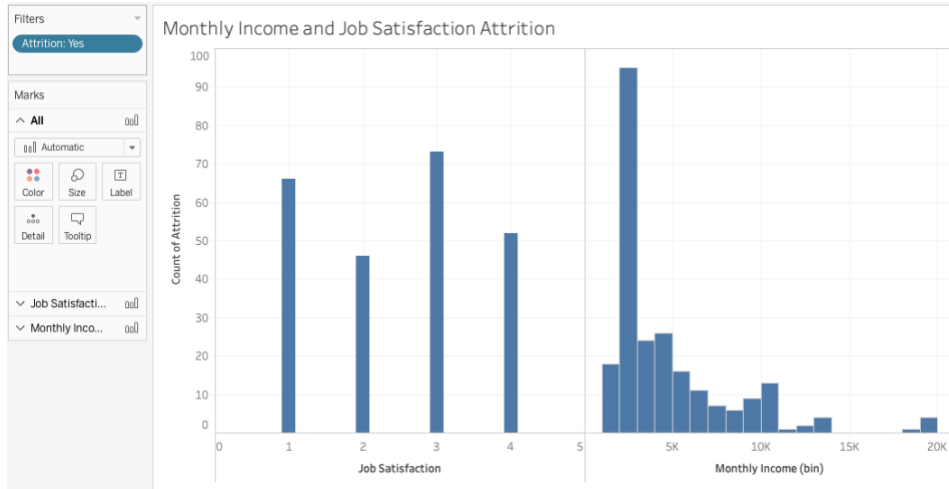
Attrition by Age & Gender



Analysis Summary: In the above graphs shows, I tried to dig further to find out in which age group, the attrition rate is higher for male and female. The visual graph above shows that except the age group 20-24, male attrition rate is higher than that of female

Key takeaway: Attrition rate is higher for male than female in most of the age groups except the age group 20-24.

Attrition by Monthly income & Job satisfaction



Analysis Summary: After analyzing the impact of “monthly income” on attrition, I found that as the monthly income increasing, attrition rate is decreasing. I tried to dig deeper to find out if employees with lower monthly income has lower job satisfaction that is causing higher attrition rate in that group. The above two graphs are showing that as the monthly income is increasing, job satisfaction level is increasing, and the attrition rate is decreasing and vice-versa. The above analysis shows that job satisfaction level is low with low monthly income that can be the reason of high attrition rate with low monthly income.

Key takeaway: With high monthly income, job satisfaction level is high, and the attrition rate is low.

Summary of the Exploratory analysis

Attributes	Takeaway from Analysis
<i>Age</i>	Negatively related to attrition
<i>Gender</i>	Attrition rate between male and female is differing by almost 2%
<i>Marital status</i>	Attrition rate of single is almost two times the attrition rate of married and divorced
<i>Number of companies worked</i>	Attrition rate is overall positively related to number of companies worked
<i>Total working years</i>	Attrition rate is reversely related to total working years of employees
<i>Years at company</i>	Attrition rate is reversely related to years at company
<i>Years since last promotion</i>	Attrition rate is reversely related to “years-since-last-promotion”
<i>Total number of Years with his/her current manager</i>	Attrition rates are contrariwise related to years with current manager of the employee
<i>Job involvement</i>	Attrition rate is going down with the increase in the job involvement levels.
<i>Environment satisfaction</i>	Attrition rate is going down with the increase in the environment satisfaction level
<i>Relationship satisfaction</i>	Attrition rate is overall going down with the increase in relationship satisfaction level
<i>Monthly income</i>	Downward trend in the attrition rate with respect to increase in monthly income
<i>Overtime</i>	Attrition rate is higher for employees who works overtime
<i>Education field</i>	Attrition rate is higher at human resources and technical degree as compared to other education fields
<i>Department</i>	Sales department has the most attrition followed by Human Resource and Research and Development
<i>Job level</i>	Attrition rate is inversely related to job level
<i>Business travel</i>	Attrition rate is higher for employees who do business travel frequently
<i>Distance from home</i>	As the distance from office increases, employees are more likely to quit their job.

<i>Department & Job role</i>	Sales representative, Sales executives and laboratory technicians have the highest attrition rates due to lower remuneration for these roles.
<i>Age & Gender</i>	Attrition rate is higher for male than female in most age groups except the age group of 20-24
<i>Monthly income & Job satisfaction</i>	With high monthly income, job satisfaction level is high, and the attrition rate is low

Predicting Employee Attrition at the organization using ML models

In this chapter, I did the analysis to predict attrition rate based on various features available in the dataset. There are various classifier models available in Python's scikit-learn library. I am using the Logistic Regression linear classification model to run predictive analysis and the goal of the model is to predict the chances of an employee leaving the company given their individual circumstances i.e. employee, demographics, performance ratings, compensation details etc.

I choose logistic regression as I am looking for a solution for a classification problem. The classification problem in my term paper is "Attrition" that has binary outcome either employees are quitting or not quitting. Logistic regression is a popular classification technique to predict binary outcomes ($y = 0$ or 1) (Nagesh Singh Chauhan, 2019). Logistic regression will give the probability of an event occurrence, in this my case chance of employees quitting based on various attributes that I am going to consider in my model. The attributes that I am going

I will keep the continuous variable as it is. However, for category variables I will create dummy variables by using python libraries and will create new columns for each of the category values. In order to run logistic regression model, I am using python. In python, I will then split the data into training set and testing set in the 80/20 ratio. I will create an instance of the model with all the parameters in the training model and the idea is to check what parameters are significant in predicting employee attrition. I will iterate on the model and will eliminate attributes that are not valuable in predicting attrition rate in each iteration till I am only left with only those features that are valuable in predicting attrition. I will then test the accuracy of the model by predicting attrition for the test data set. To check the performance of my model, I will use score method in python to arrive at the accuracy rate. I will also check the ROC curve

and the confusion matrix. Confusion matrix will provide me the summary of prediction with breaking down into “number of correct and incorrect prediction” (Confusion Matrix in Machine Learning, 2020). In order to run the confusion matrix, I will use either of the two Seaborn and Matplotlib packages from python.

In order to predict employee attrition at the organization, here are steps that I followed to fine tune and run the Logistic Regression linear classifier and Random Forest tree classifier

Step 1: Importing the dataset and relevant libraries in Python

Importing relevant packages in Python

```
import pandas as pd
import numpy as np

import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import LabelEncoder

from sklearn.model_selection import train_test_split
from sklearn.model_selection import GridSearchCV
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score, confusion_matrix, precision_score, recall_score, f1_score
```

Setting options to optimize the print function output in Python.

```
sns.set_style("whitegrid")
plt.style.use("fivethirtyeight")

'''Set the options to extend the width of the print function output and the column output'''
desired_width=700
pd.set_option('display.width', desired_width)
np.set_printoptions(linewidth=desired_width)
pd.set_option('display.max_columns',34)
```

Reading the dataset into a dataframe object from pandas library.

```
hrAnalyticsDataFrame = pd.read_csv("WA_Fn-UseC_-HR-Employee-Attrition.csv")
```

Describing the dataset and printing the output

```
print(hrAnalyticsDataFrame.describe())
```

Prepare and validate the dataset

1. Validate that there are no null or missing values in the data.

```
"""Check how many missing values/null fields"""
print("\nMissing values: ", hrAnalyticsDataFrame.isnull().sum().values.sum())

Missing values:  0
```

The dataset is clean and has no missing values.

2. Check the validity of the dataset by making sure that the following validations are true for each observation.

- a. TotalWorkingYears >= YearsAtCompany
- b. YearsAtCompany >= YearsInCurrentRole
- c. YearsAtCompany >= YearsSinceLastPromotion
- d. YearsAtCompany >= YearsWithCurrManager
- e. MonthlyRate >= DailyRate
- f. DailyRate >= HourlyRate

```
def compare_column_values(firstcolumnname, secondcolumnname, hrData):
    for indexofCol, valueofCol in hrData[firstcolumnname].iteritems():
        if valueofCol < hrData[secondcolumnname][indexofCol]:
            print(f"The validity that {firstcolumnname} >= {secondcolumnname} is NOT TRUE")
            break
    print(f"The validity of {firstcolumnname} >= {secondcolumnname} is TRUE")

"""Validate that TotalWorkingYears >= YearsAtCompany"""
compare_column_values('TotalWorkingYears', 'YearsAtCompany', hrAnalyticsDataFrame)

"""Validate that YearsAtCompany >= YearsInCurrentRole"""
compare_column_values('YearsAtCompany', 'YearsInCurrentRole', hrAnalyticsDataFrame)
```

```

"""Validate that YearsAtCompany >= YearsSinceLastPromotion"""
compare_column_values('YearsAtCompany', 'YearsSinceLastPromotion', hrAnalyticsDataFrame)

"""Validate that YearsAtCompany >= YearsWithCurrManager"""
compare_column_values('YearsAtCompany', 'YearsWithCurrManager', hrAnalyticsDataFrame)

"""Validate that MonthlyRate >= DailyRate"""
compare_column_values('MonthlyRate', 'DailyRate', hrAnalyticsDataFrame)

"""Validate that DailyRate >= HourlyRate"""
compare_column_values('DailyRate', 'HourlyRate', hrAnalyticsDataFrame)

```

3. Analyze the dataset to find any columns that have a standard deviation of 0. Such columns will not have any predictive value since they either have the same value for all observations in the dataset. The following columns were found to meet this criterion and were removed from the dataframe.

- a. EmployeeCount
- b. Over18
- c. StandardHours

Also, identify any columns that have unique value for every observation in the dataset.

We found that the EmployeeNumber column satisfied this condition and hence was removed from the dataframe.

```

"""Identify any columns that have 0 standard deviation"""
for column in hrAnalyticsDataFrame.columns:
    if hrAnalyticsDataFrame[column].nunique() == 1 or hrAnalyticsDataFrame[column].nunique() == 1470:
        print(f"{column}: Number of unique values {hrAnalyticsDataFrame[column].nunique()}")

"""Identify that have unique values for all the rows"""

```

```

for column in hrAnalyticsDataFrame.columns:
    if hrAnalyticsDataFrame[column].nunique() == 1 or hrAnalyticsDataFrame[column].nunique() == 1470:
        print(f"{column}: Number of unique values {hrAnalyticsDataFrame[column].nunique()}")

hrAnalyticsDataFrame.drop(['EmployeeCount', 'EmployeeNumber', 'Over18', 'StandardHours'],
axis="columns", inplace=True)

hrAnalyticsDataFrame.drop(['EmployeeCount', 'EmployeeNumber', 'Over18', 'StandardHours'])

```

Step 2: Data processing – Encoding category variables and feature selection

In this section, I was trying to understand if the data frame has all the relevant category variables correctly identified. If not, then the idea is to transform them into category variables.

Additionally, I will try to find if features are closely related to each other (multicollinearity) which will lead to incorrect results from the Logistic regression model. For highly correlated features, the idea is to drop one of the features that has lower feature importance for predicting attrition and remove collinearity from the dataset.

1. Analyze the dataset to find out if all the category variables have been correctly identified.

```

for column in hrAnalyticsDataFrame.select_dtypes(['object']).columns:
    columns_Of_Object_Type.append(column)
    print(f"{column} : {hrAnalyticsDataFrame[column].unique()}")

```

Initial analysis of the dataframe in python identified that the following columns had the incorrect datatype assigned and are not identified as category variables.

- a) Attrition
- b) BusinessTravel
- c) Department

- d) EducationField
- e) Gender
- f) JobRole
- g) MaritalStatus
- h) OverTime

2. Encode the variables identified in step 4 to convert them into dummy variables.

```
categorical_data = {  
    'Attrition': {'No': 0, 'Yes': 1},  
    'BusinessTravel': {'Non-Travel': 0, 'Travel_Frequently': 1, 'Travel_Rarely': 2},  
    'Department': {'Human Resources': 0, 'Research & Development': 1, 'Sales': 2},  
    'EducationField': {'Human Resources': 0, 'Life Sciences': 1, 'Marketing': 2, 'Medical': 3, 'Technical Degree':  
4,  
        'Other': 5},  
    'Gender': {'Female': 0, 'Male': 1},  
    'JobRole': {'Healthcare Representative': 0, 'Human Resources': 1, 'Laboratory Technician': 2, 'Manager': 3,  
        'Manufacturing Director': 4, 'Research Director': 5, 'Research Scientist': 6, 'Sales Executive': 7,  
        'Sales Representative': 8},  
    'MaritalStatus': {'Divorced': 0, 'Single': 1, 'Married': 2},  
    'OverTime': {'No': 0, 'Yes': 1}}  
  
hrAnalyticsDataFrame = hrAnalyticsDataFrame.replace(categorical_data)
```

3. Check for multicollinearity among the features in the dataset. The idea is to find the correlation between the features in the dataset by using the heatmap function in python. If the correlation between features other than attrition is > 0.5 then there is multicollinearity between the features and one of them has to be removed.

```
features_with_high_correlation = hrAnalyticsDataFrame.corr().abs().nlargest(20, "Attrition").Attrition.index  
  
# Plotting a diagonal correlation matrix.  
# Referred from # https://seaborn.pydata.org/examples/many\_pairwise\_correlations.html  
sns.set_style("white")
```


From the above correlation matrix, it is clear that multicollinearity exists between the following features:

- a. “Monthly Income” is highly correlated with “Job Level” (0.95), “Total Working Years” (0.77), “Years at company” (0.51) and “Age” (0.50). Since, the hypothesis is that “Monthly Income” is more important to an employee than all of these features, the decision is to drop all of these features from the dataset and retain “Monthly Income”.
- b. “Department” and “Job Role” are highly correlated (0.66). Since, the hypothesis is that “Job Role” is more important to an employee than “Department”, the decision is to drop “Job Level” from the dataset and retain “Department”.
- c. “Years at company” and is highly correlated with “Years with current manager” (0.77) and “Years in Current Role” (0.76). Since, “Years at company” was removed since it was correlated to “Monthly Income”, I will remove both “Years with current manager” and “Years in Current Role” from the dataset.

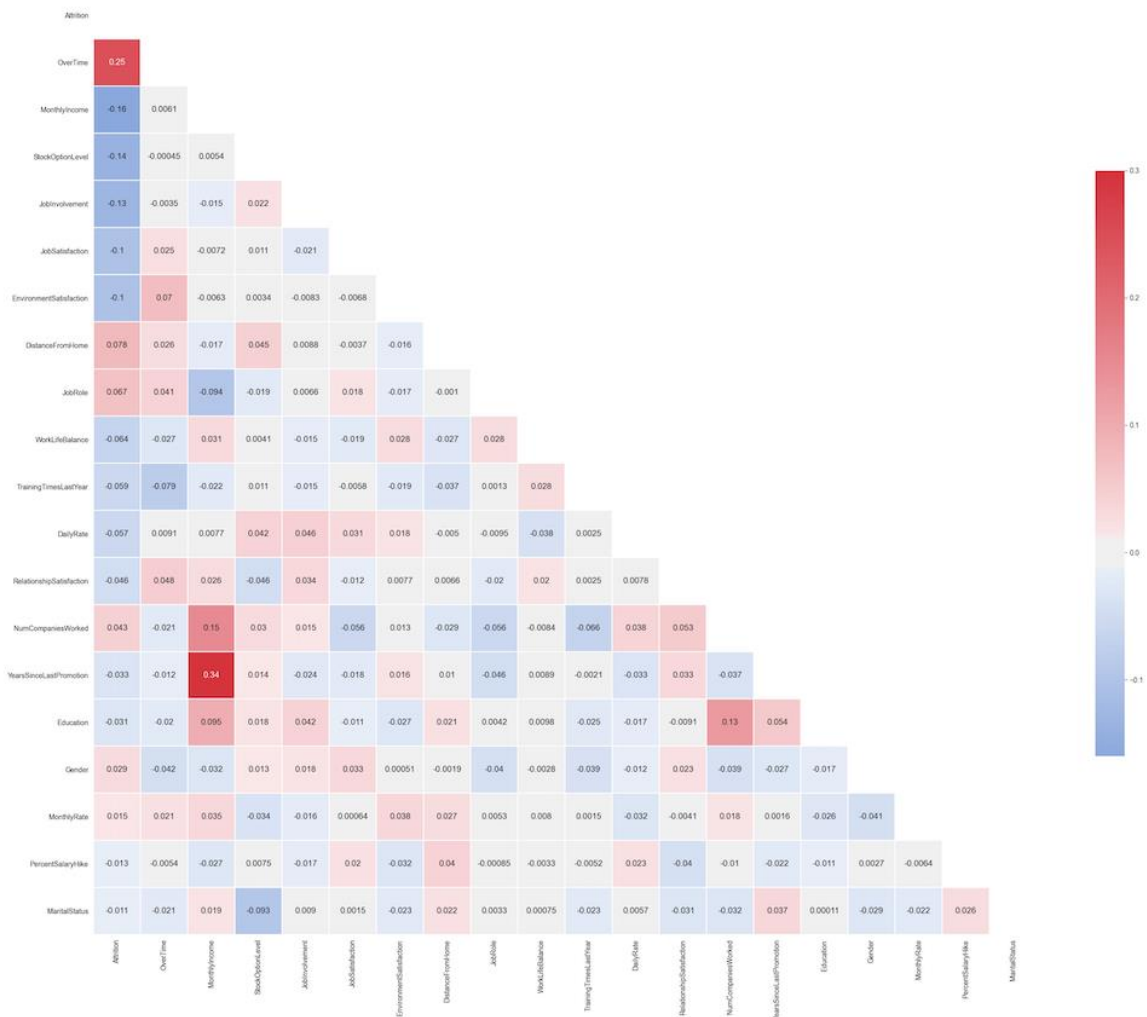
```
hrAnalyticsDataFrame = hrAnalyticsDataFrame.drop(columns=['JobLevel'])
hrAnalyticsDataFrame = hrAnalyticsDataFrame.drop(columns=['TotalWorkingYears'])
hrAnalyticsDataFrame = hrAnalyticsDataFrame.drop(columns=['YearsAtCompany'])
hrAnalyticsDataFrame = hrAnalyticsDataFrame.drop(columns=['Department'])
hrAnalyticsDataFrame = hrAnalyticsDataFrame.drop(columns=['Age'])
hrAnalyticsDataFrame = hrAnalyticsDataFrame.drop(columns=['YearsInCurrentRole',
'YearsWithCurrManager'])

fig, ax = plt.subplots(figsize=(30, 30))

features_with_high_correlation = hrAnalyticsDataFrame.corr().abs().nlargest(20, "Attrition").Attrition.index
sns.heatmap(hrAnalyticsDataFrame[features_with_high_correlation].corr(), mask=mask, cmap=cmap,
vmax=.3, center=0, annot=True,
          annot_kws={"size": 12}, square=True, linewidths=.5, cbar_kws={"shrink": .5})
```

```
plt.savefig("CorrelationMatrix2")
print(hrAnalyticsDataFrame.info())
```

The resulting correlation matrix after removing all features that were causing the multicollinearity in the dataset is shown below. It is clear that the problem of multicollinearity in the dataset has been resolved and the dataset will not have incorrect or biased results due to processing using linear classification models.



Step 3: Split the dataset into training and test datasets

1. In order to run the model, we will first specify the feature variables and the output variable as shown below

```
X = hrAnalyticsDataFrame.drop("Attrition", axis=1)
y = hrAnalyticsDataFrame.Attrition
```

2. The idea here is to split the dataset into training and test datasets. The training data will consist of 80% of the data and the test set will consist of the remaining 20%. We will use the training dataset to train the model and then use the test dataset to measure the performance of the model. In the output dataset, it is important to maintain the proportion of "Attrition" values of "0" (did not quit their job) and "1" (quit their job) in the ratio of 84%:16% as present in the original dataset. The original dataset contains employee records where 16% of the employee quit their jobs and 84% did not quit their jobs. It is important to maintain the same proportion in both the test and training datasets.

```
# Stratify the dataset to maintain the 84:16 ratio of the Attrition variable in the test and training sets.
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42, stratify=y)
kf = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)
```

3. The next step is to initialize the Logistic Regression model from the sklearn.linear_model package and call the fit method to train the model with default parameters.

```
logisticRegressionModel = LogisticRegression(C=1.0, random_state=42, multi_class="auto", n_jobs=-1,
max_iter=10000)
logisticRegressionModel.fit(X_train, y_train)
```

4. Now, in order to evaluate the performance of the model, we will look at the accuracy, precision, recall and F1 scores for the model. In order to view the scores of this model for both the training and test datasets, I have created the following function.

```
# Function to print the accuracy, precision, recall and F1 scores for the logistic regression model
def print_model_score(model, X_train, y_train, X_test, y_test, optimal_threshold=0.5, train=True):
    if train:
        predict_attrition = (model.predict_proba(X_train)[:, 1] >= optimal_threshold)
        print("Train Result:\n=====")
        print(f"Accuracy score: {accuracy_score(y_train, predict_attrition):.4f}\n")
        print(f"Logistic Score Report: \n \tPrecision: {precision_score(y_train, predict_attrition)}\n\t"
              f"Recall Score: {recall_score(y_train, predict_attrition)}\n\t"
              f"F1 score: {f1_score(y_train, predict_attrition)}\n")
        print(f"Confusion Matrix: \n {confusion_matrix(y_train, predict_attrition)}\n")
    else:
        predict_attrition = (model.predict_proba(X_test)[:, 1] >= optimal_threshold)
        print("Test Result:\n=====")
        print(f"Accuracy score: {accuracy_score(y_test, predict_attrition)}\n")
        print(f"Logistic Score Report: \n \tPrecision: {precision_score(y_test, predict_attrition)}\n\t"
              f"Recall Score: {recall_score(y_test, predict_attrition)}\n\t"
              f"F1 score: {f1_score(y_test, predict_attrition)}\n")
        print(f"Confusion Matrix: \n {confusion_matrix(y_test, predict_attrition)}\n")
```

For the first iteration, I have assumed the threshold value for classifying predicted value from the model to be 0.5. Hence, a value ≥ 0.5 from the model is interpreted as a prediction of “attrition” and a value ≤ 0.5 is interpreted as a prediction of “No attrition”. With a Threshold value of 0.5, following are the scores for the model.

```
print_model_score(logisticRegressionModel, X_train, y_train, X_test, y_test, optimal_threshold=0.5,
train=False)
```

Test Result:

=====

Accuracy score: 0.8775510204081632

Logistic Score Report:

Precision: 0.9230769230769231

Recall Score: 0.2553191489361702

F1 score: 0.4

Confusion Matrix:

```
[[246  1]
 [ 35 12]]
```

The baseline model score is terrible with an F1 score of 40%. The model does a poor job in predicting attrition i.e. employees who are likely to quit and only has an accuracy of 25% in predicting attrition for an employee from the test dataset.

5. The next step was to optimize the model by finding the optimal value for the hyperparameter C. The idea here is to find the optimal value of the F1 score for the model by iterating through all the possible C values between 0.001 and 1000 and using StratifiedKFold to select the training and test datasets.

```
kf = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)
f1_score_list = []

# Create a list of C values varying from 0.001 to 1000, using StratifiedKFold cross validation to train data
c_list = 10**np.linspace(-3, 3, 200)

for c in c_list:
    logisticRegressionModel = LogisticRegression(C=c)
    cvs = (cross_val_score(logisticRegressionModel, X_train, y_train, cv=kf, scoring='f1')).mean()
    f1_score_list.append(cvs)
optimal_c = float(c_list[f1_score_list.index(max(f1_score_list))])
print('optimal value of C = {:.3f}'.format(optimal_c))
```

The optimal C for the model is determined to be 0.420. On running the model with the default threshold value of 0.5, we see a deterioration in the performance of the model with the F1 score dropping to 29%.

```
logisticRegressionModel = LogisticRegression(C=optimal_c, max_iter=10000)
logisticRegressionModel.fit(X_train, y_train)
```

```
print_model_score(logisticRegressionModel, X_train, y_train, X_test, y_test, optimal_threshold=0.5,
train=True)
print_model_score(logisticRegressionModel, X_train, y_train, X_test, y_test, optimal_threshold=0.5,
train=False)
```

Test Result:

```
=====
```

Accuracy score: 0.8707482993197279

Logistic Score Report:

Precision: 0.8461538461538461

Recall Score: 0.23404255319148937

F1 score: 0.36666666666666675

Confusion Matrix:

```
[[245  2]
```

```
[ 36 11]]
```

6. It is clear from the previous step that there is a need to optimize the threshold value to improve the prediction power of the model. In order to optimize the threshold value for the model, I started iterating through values between 0.1 and 0.9 to determine if the F1 is improving near a particular threshold value. At the end of this iteration, it was clear that the optimal threshold value for the model is near 0.3. The model scores at a threshold value of 0.3 as shown below:

```
print_model_score(logisticRegressionModel, X_train, y_train, X_test, y_test, optimal_threshold=0.3,
train=True)
print_model_score(logisticRegressionModel, X_train, y_train, X_test, y_test, optimal_threshold=0.3,
train=False)
```

Test Result:

```
=====
```

```
Accuracy score: 0.8707482993197279
```

```
Logistic Score Report:
```

```
Precision: 0.6363636363636364
```

```
Recall Score: 0.44680851063829785
```

```
F1 score: 0.5249999999999999
```

```
Confusion Matrix:
```

```
[[235 12]
```

```
[ 26 21]]
```

We see that there is a drastic improvement in the performance of the model at a threshold value of 0.3 with the F1 score improving to 53% and the recall score improving to 44.7%.

Conclusion

The model performance could be optimized to reach an F1 score of 53%, a recall score of 44.7% and precision score of 63.6%. As part of optimizing the model, it is to be noted that the precision score went down from 92% (when no optimizations were applied) to 63.6%. However, it is to be noted that it is more valuable for the organization at being able to predict which employees are likely to quit their jobs as compared to the downside of an increase in the number of false positive i.e. incorrectly predicting that certain employees are likely to leave the organization. Hence, the threshold value was optimized to improve the recall score while taking a hit on the precision score i.e. optimizing true positives and false negatives at the expense of false positives.

Opportunities to improve model performance

However, the model score is still not very good and the following steps can be taken to improve the prediction of attrition at the organization.

1. Feature engineering can be utilized to create new derived from the existing feature set which might lead to a better fitting model with greater accuracy in predicting Attrition.

2. The dataset has a class imbalance problem i.e. it is the problem in machine learning where the total number of a class of data (positive) is far less than the total number of another class of data (negative) (Devin Soni, 2018). In the dataset, 84% of the data is about employees not quitting their jobs as against 16% of the data representing employees quitting their jobs. The model performance can be improved by applying Imbalance Class mitigation techniques such as SMOTE sampling, Cost-sensitive learning and Anomaly detection (Devin Soni, 2018).
3. Other classification models can be used to optimize the performance of predicting attrition at the organization for example; tree-based classification models such as Random Forest, Decision Tree Classifier; Bayesian classification models Gaussian Naive Bayes, Bernoulli Naive Bayes, Multinomial Naive Bayes and others.
4. Last but not the least, additional observations will greatly help in improving the performance of the model by including data about more employees and tracking the same information. A larger dataset will help improve the training of the model thereby improving its performance.

Prescriptive Analytics to minimize attrition and retain top performers

In this chapter, my aim will be to find out the measures that a company can take to avoid attrition. In terms of prescriptive analytics, the idea is to determine the top 10 features that are important in predicting attrition and track those features closely to determine the factors on which to focus and improve the attrition rate at the organization.

In order to determine the top 10 key features that led to employees quitting their jobs in the past at the organization, I ran a random forest classifier which provides the list of important features that are important in predicting attrition using the below code in python.

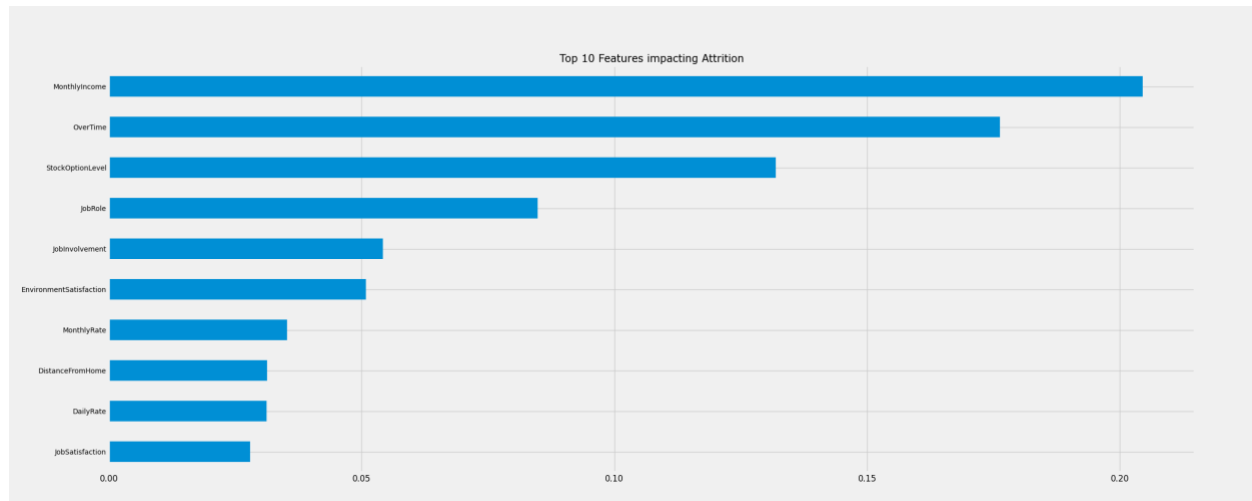
```
n = 10 # specify n (top n features)
randomforestclassifier = RandomForestClassifier(random_state=42, n_estimators=200, max_depth=3)
randomforestclassifier_predictions = randomforestclassifier.fit(X, y)

# Sort the feature in their order of importance in predicting attrition
pd.Series(randomforestclassifier_predictions.feature_importances_, index=X.columns).nlargest(n).plot(kind='barh',
                                                    figsize=[25, n / 2]
                                                    ).invert_yaxis()

ticks_x = np.linspace(0, 0.2, 5) # (start, end, number of ticks)
plt.xticks(ticks_x, fontsize=12, color='black')
plt.yticks(size=10, color='Orange')
plt.title('Top 10 Features impacting Attrition', family='Verdana', size=15)

plt.savefig("Top 10 features")
```

The below figure lists the top 10 features that have led to attrition at the organization.



Key Recommendations

1. The model clearly shows that the most important features at predicting attrition are monetary attributes such as overtime, monthly salary and daily rate. Hence, it is important to focus on the remuneration structure at the organization and determine if there are improvements necessary to improve employee satisfaction and bring compensation to industry standards.
2. Secondly, job satisfaction and company culture are important for employees since both Job Satisfaction, Job role and Environment Satisfaction are in the top 10 list. Hence, it will be important for HR to focus on employee motivation and engagement and improve company culture to keep employees engaged.
3. Additionally, the distance of the office from home is important to employees and it will be useful to keep that in mind when recruiting new employees at the organizations.
4. Another key finding from exploratory data analysis was that performance rating data was pretty restrictive in the dataset with only values of 3 and 4. It leads to a hypothesis that the performance rating system at the organization is not very fair and does not recognize top performers at the organization. HR might need to look at the performance rating

system and do an overhaul that should lead to improved job satisfaction and engagement among employees.

The above recommendations can be used to derive guidelines around what sort of employees should be hired in the future and improve the culture and remuneration within the organization so that the attrition rate is low in the future at the organization.

Conclusion

Employees are a crucial part of the organization. The way the organization performs is dependent heavily on the performance of its employees. To achieve efficient and effective performance from employees, an organization needs to consider certain factors in order to retain such key employees. This term paper also shows the importance of HR analytics and how it can be used to bring down hiring costs at an organization through digital transformation. The term paper also provides insights that can help HR take decisive steps towards improving organization processes so that employee performance attrition can be lowered in the future.

References

- Al-Habil, W. I., Allah, A., & Shehadah, M. (2017, November 28). Factors Affecting the Employees' Turnover at the Ministry of High Education in Gaza Governorates-Case study: North and West Gaza Directorates of Education. Retrieved from <https://www.omicsonline.org/open-access/factors-affecting-the-employees-turnover-at-the-ministry-of-high-education-in-gaza-governoratescase-study-north-and-west-gaza-dire-2151-6200-1000304-95871.html>
- Confusion Matrix in Machine Learning. (2020, February 23). Retrieved from <https://www.geeksforgeeks.org/confusion-matrix-machine-learning/>
- Chamberlain, A. (2018, February 28). Why Do Employees Stay? A Clear Career Path and Good Pay, for Starters. Retrieved from <https://hbr.org/2017/03/why-do-employees-stay-a-clear-career-path-and-good-pay-for-starters>
- Chauhan, N. S. (2019, April 4). Real world implementation of Logistic Regression. Retrieved from <https://towardsdatascience.com/real-world-implementation-of-logistic-regression-5136cefb8125>
- Galarnyk, M. (2020, February 13). Logistic Regression using Python (scikit-learn). Retrieved from <https://towardsdatascience.com/logistic-regression-using-python-sklearn-numpy-mnist-handwriting-recognition-matplotlib-a6b31e2b166a>

Habil, W. I. A., Allah, A., & Shehadah, M. (2017). Factors Affecting the Employees' Turnover at the Ministry of High Education in Gaza Governorates-Case study: North and West Gaza Directorates of Education. *Arts and Social Sciences Journal*, 08(05). doi: 10.4172/2151-6200.1000304

Mamun, C. A. A., & Hasan, M. N. (2017). Factors affecting employee turnover and sound retention strategies in business organization: a conceptual view. *Problems and Perspectives in Management*, 15(1), 63–71. doi: 10.21511/ppm.15(1).2017.06

Mishra, S. (n.d.). REVIEW OF LITERATURE ON FACTORS INFLUENCING ATTRITION AND RETENTION. Retrieved from https://www.academia.edu/39006530/REVIEW_OF_LITERATURE_ON_FACTORS_INFLUENCING_ATTRITION_AND_RETENTION

Rogers, M. (2020, January 20). A Better Way to Develop and Retain Top Talent. Retrieved from <https://hbr.org/2020/01/a-better-way-to-develop-and-retain-top-talent>

“API Reference¶.” *Scikit*, scikit-learn.org/stable/modules/classes.html.

“Sklearn.model_selection.cross_validate” *Scikit*, scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_validate.html.

sklearn.linear_model.LogisticRegression. (n.d.). Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

Soni, D. (2019, July 16). Dealing with Imbalanced Classes in Machine Learning. Retrieved from <https://towardsdatascience.com/dealing-with-imbalanced-classes-in-machine-learning-d43d6fa19d2>

Using Seaborn Python Package for Creating Heatmap. (2020, March 3). Retrieved from <https://blog.quantinsti.com/creating-heatmap-using-python-seaborn/>