"I Quit"

Analysis of Predictive Models for Employee Attrition

Lok H Ngan, Andrew Narbutis, Justin Pate | IST 707 | June 14, 2019

THE iSCHOOL Syracuse University

1. Introduction

Background

"The simplest way to stop your employees from leaving is to develop a plan to make them stay."

Anonymous

In both April and May 2019, the Bureau of Labor Statistics reported that the unemployment rate had fallen to a 50 year low of 3.6% (figure 1) ("Employment Situation Summary", 2019). With many employers trying to attract employees to support growing operations, that low rate means that employers seeking to fill open positions must choose from either the people who are unemployed and looking for work or attract workers already employed with other companies. The same BLS report shows that the rate of employees voluntarily quitting their current positions continues to grow, more than doubling in the last 10 years to a current rate of about 42 million per year (figure 2) ("Employment Situation Summary", 2019).

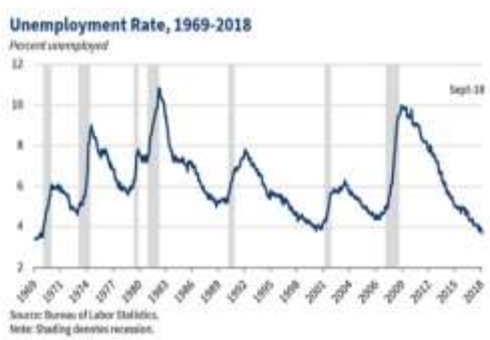


Figure 1: U.S. Unemployment Rate 1969-2018

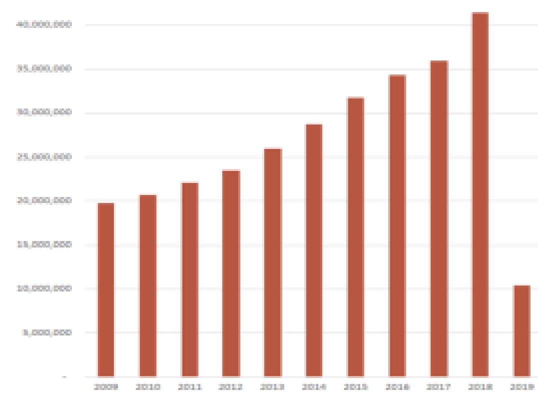


Figure 2: Annual voluntary departures

These figures underscore the current employment atmosphere as a buyer's market. In other words, workers are in high demand, there are not enough workers to meet the needs of business leaders, and the competition for those workers continues to grow. Workers, now facing the ability to select from several competing offers are willing to leave their current employers in pursuit of a better job at a better company for better compensation. The term "war for talent" was coined by McKinsey's Steven Hankin in 1997 and popularized by the book of that name in 2001 (Axelrod, Handfield-Jones, Michaels, 2001). That war is being fought harder than ever in this current environment.

Employees are both an investment and an appreciating asset. They are an investment because of all the time and effort it takes to find, train and develop them. And they appreciate in value because over time they become more efficient at their work, they can identify valuable relationships between business processes and identify ways to eliminate waste. With almost every passing day, an employee learns more about the company and their own job and can turn around and apply that knowledge to achieve more the next day. While that value can grow in different ways and at different rates for each unique situation, the increase in value is nearly universal.

Like other valuable assets, companies want to keep most employees. The cost of employee turnover is high and can range anywhere from a few thousand dollars to estimates as high as 1.5-2x an employee's annual salary. Consider the following tangible and intangible costs that are associated with turnover:

- Time spent to complete exit interviews and termination processing,
- Costs (real dollars and time) to advertise open positions, interview, hire perform background checks, orientation, training, and paying referral bonuses
- Until an open position is filled, the work normally performed by that employee needs to be covered by other employees, diverting from work or increasing overtime
- The costs of work that simply is not completed that hurts the top line (e.g. sales)
- The time it takes for a new hire to come up to speed and become efficient
- Time required by manager and coworkers to bring a new hire up to speed
- Lost institutional knowledge

The internet and professional publications are full of different points of view supporting different models on how to value the cost of turnover (e.g. Bersin, 2013; Fortin, 2017). The one constant and accepted norm among all these models is that employee turnover does have a negative impact to a company's finances and productivity (Morrell, 2014).

In this competitive atmosphere, many companies are turning to the data they maintain about their employees to try to determine what will entice their employees to stay. Although most companies keep track of employee turnover, many fall short when they try to understand its causes and costs in a meaningful way. This paper attempts to show some strategies on how to take a data driven approach to understanding the drivers of turnover as well as provide some recommendations for actions that can be taken to address turnover based on those results.

Business Question

The primary business question that is being asked is "Can HR data be used to understand the drivers of turnover within an organization?"

The remainder of this paper will focus on answering that question.

2. Data Overview

Data Source

Human Resources data is extremely sensitive and protected by a number of privacy and data protection related laws. For this reason, it is extremely difficult to locate a comprehensive HR dataset to use for this sort of analysis.

The data set used in this analysis was created by the IBM Watson team. This dataset has a total of 1470 observations with 28 variables for each observation. To be clear, it is not a dataset with employee data from an existing company. The IBM website states "This is a fictional data set created by IBM data scientists. Its main purpose is to demonstrate the Watson analytics tool for employee attrition." Therefore, the dataset is useful for creating and testing models with data that is like data that would be maintained by a large company but should not be considered real world set of data. The dataset can be accessed via the following link:

https://www.ibm.com/communities/analytics/watson-analytics-blog/hr-employee-attrition/

It is suggested that the techniques discussed in this paper be leveraged by HR analytics practitioners using data provided by their employers.

Data Quality

The dataset is complete. There are no missing values and every vector is complete. This was probably done by the IBM authors to facilitate the focus on building and testing models with the data rather than focusing on the equally important aspect of cleaning and preparing the data for analysis.

Data Dictionary (raw state, not factorized or discretized)

Variable	Description
Age	Numerical Value
Attrition	Employee leaving the company (0=no, 1=yes)
BusinessTravel	(1=No Travel, 2=Travel Frequently, 3=Travel Rarely)
DailyRate	Numerical Value
Department	(1=HR, 2=Research & Development, 3=Sales)
DistanceFromHome	Numerical Value – The distance in miles from home to employee's work location
Education	Numerical Value (1 = High school, 2 = some college, 3 = Bachelors, 4 = Masters, 5 = PhD+
EducationField	(1=HR, 2=Life Sciences, 3=Marketing, 4=Medical Sciences, 5=Other, 6= Technical)
EmployeeCount	Numerical Value
EmployeeNumber	Numerical Value - Employee ID
EnvironmentSatisfaction	Numerical Value – meaning/scale unknown
Gender	(1=Female, 2=Male)
HourlyRate	Numerical Value – Hourly rate of pay or equivalent
JobInvolvement	Numerical Value – meaning/scale unknown
JobLevel	Numerical Value – Hierarchical Job Level (relative: 1 = lowest, 5 = highest)
JobRole	(1 = Healthcare Rep, 2 = HR, 3 = Lab Technician, 4 = Manager, 5 = Managing Director, 6 = Research Director, 7 = Research Scientist, 8 = Sales Executive, 9 = Sales Representative)
JobSatisfaction	Numerical Value - (relative: 1 = lowest, 4 = highest)
MaritalStatus	(1 = Divorced, 2 = Married, 3 = Single)
MonthlyIncome	Numerical Value – Monthly Salary
MonthlyRate	Numerical Value - meaning/scale unknown
NumCompaniesWorked	Numerical Value – Number of companies the employee has worked at
Over18	(1 = Yes, 2 = No)
OverTime	(1=No, 2=Yes)
PercentSalaryHike	Numerical Value – Percentage increase in Salary at last review The percentage of change in salary between last two years
PerformanceRating	Numerical Value – Performance Rating (relative: 1 = lowest, 4 = highest)
RelationshipSatisfaction	Numerical Value - (relative: 1 = lowest, 4 = highest)
StandardHours	Numerical Value – Standard hours worked per 2 week period
StockOptionLevel	Numerical Value – Level for determining amount of stock options offered (relative: 1 = lowest, 4 = highest)
TotalWorkingYears	Numerical Value – Total number of years working (all companies)

TrainingTimesLastYear	Numerical Value – Number of training sessions attended in the last year
WorkLifeBalance	Numerical Value – Quality of work life balance (relative: 1 = lowest, 4 = highest)
YearsAtCompany	Numerical Value – Total number of years working for this company
YearsInCurrentRole	Numerical Value -Number of years in current role
YearsSinceLastPromotion	Numerical Value – Number of years since last promotion
YearsWithCurrManager	Numerical Value – Number of years spent working for current manager

<u>Data Summary</u>

```
##       Age         Attrition          BusinessTravel       DailyRate     
##  Min.   :18.00   Length:1470        Length:1470        Min.   : 102.0  
##  1st Qu.:30.00   Class :character   Class :character   1st Qu.: 465.0  
##  Median :36.00   Mode  :character   Mode  :character   Median : 802.0  
##  Mean   :36.92                                         Mean   : 802.5  
##  3rd Qu.:43.00                                         3rd Qu.:1157.0  
##  Max.   :60.00                                         Max.   :1499.0  

##   Department        DistanceFromHome   Education     EducationField    
##  Length:1470        Min.   : 1.000   Min.   :1.000   Length:1470       
##  Class :character   1st Qu.: 2.000   1st Qu.:2.000   Class :character  
##  Mode  :character   Median : 7.000   Median :3.000   Mode  :character  
##                     Mean   : 9.193   Mean   :2.913                     
##                     3rd Qu.:14.000   3rd Qu.:4.000                     
##                     Max.   :29.000   Max.   :5.000                     

##  EmployeeCount EmployeeNumber   EnvironmentSatisfaction    Gender         
##  Min.   :1     Min.   :   1.0   Min.   :1.000           Length:1470       
##  1st Qu.:1     1st Qu.: 491.2   1st Qu.:2.000           Class :character  
##  Median :1     Median :1020.5   Median :3.000           Mode  :character  
##  Mean   :1     Mean   :1024.9   Mean   :2.722                             
##  3rd Qu.:1     3rd Qu.:1555.8   3rd Qu.:4.000                             
##  Max.   :1     Max.   :2068.0   Max.   :4.000                             

##    HourlyRate     JobInvolvement    JobLevel       JobRole         
##  Min.   : 30.00   Min.   :1.00   Min.   :1.000   Length:1470       
##  1st Qu.: 48.00   1st Qu.:2.00   1st Qu.:1.000   Class :character  
##  Median : 66.00   Median :3.00   Median :2.000   Mode  :character  
##  Mean   : 65.89   Mean   :2.73   Mean   :2.064                     
##  3rd Qu.: 83.75   3rd Qu.:3.00   3rd Qu.:3.000                     
##  Max.   :100.00   Max.   :4.00   Max.   :5.000                     

##  JobSatisfaction MaritalStatus      MonthlyIncome    MonthlyRate   
##  Min.   :1.000   Length:1470        Min.   : 1009   Min.   : 2094  
##  1st Qu.:2.000   Class :character   1st Qu.: 2911   1st Qu.: 8047  
##  Median :3.000   Mode  :character   Median : 4919   Median :14236  
##  Mean   :2.729                      Mean   : 6503   Mean   :14313  
##  3rd Qu.:4.000                      3rd Qu.: 8379   3rd Qu.:20462  
##  Max.   :4.000                      Max.   :19999   Max.   :26999  

##  NumCompaniesWorked    Over18            OverTime        
##  Min.   :0.000      Length:1470        Length:1470       
##  1st Qu.:1.000      Class :character   Class :character  
##  Median :2.000      Mode  :character   Mode  :character  
##  Mean   :2.693                                           
```

```
##   3rd Qu.:4.000
##   Max.   :9.000

##   PercentSalaryHike PerformanceRating RelationshipSatisfaction
##   Min.   :11.00     Min.   :3.000     Min.   :1.000
##   1st Qu.:12.00     1st Qu.:3.000     1st Qu.:2.000
##   Median :14.00     Median :3.000     Median :3.000
##   Mean   :15.21     Mean   :3.154     Mean   :2.712
##   3rd Qu.:18.00     3rd Qu.:3.000     3rd Qu.:4.000
##   Max.   :25.00     Max.   :4.000     Max.   :4.000

##   StandardHours StockOptionLevel TotalWorkingYears TrainingTimesLastYear
##   Min.   :80    Min.   :0.0000   Min.   : 0.00     Min.   :0.000
##   1st Qu.:80    1st Qu.:0.0000   1st Qu.: 6.00     1st Qu.:2.000
##   Median :80    Median :1.0000   Median :10.00     Median :3.000
##   Mean   :80    Mean   :0.7939   Mean   :11.28     Mean   :2.799
##   3rd Qu.:80    3rd Qu.:1.0000   3rd Qu.:15.00     3rd Qu.:3.000
##   Max.   :80    Max.   :3.0000   Max.   :40.00     Max.   :6.000

##   WorkLifeBalance YearsAtCompany   YearsInCurrentRole
##   Min.   :1.000   Min.   : 0.000   Min.   : 0.000
##   1st Qu.:2.000   1st Qu.: 3.000   1st Qu.: 2.000
##   Median :3.000   Median : 5.000   Median : 3.000
##   Mean   :2.761   Mean   : 7.008   Mean   : 4.229
##   3rd Qu.:3.000   3rd Qu.: 9.000   3rd Qu.: 7.000
##   Max.   :4.000   Max.   :40.000   Max.   :18.000

##   YearsSinceLastPromotion YearsWithCurrManager
##   Min.   : 0.000          Min.   : 0.000
##   1st Qu.: 0.000          1st Qu.: 2.000
##   Median : 1.000          Median : 3.000
##   Mean   : 2.188          Mean   : 4.123
##   3rd Qu.: 3.000          3rd Qu.: 7.000
##   Max.   :15.000          Max.   :17.000
```

Data Structure

```
          Classes 'tbl_df', 'tbl' and 'data.frame':  1470 obs. of  37 variables:
 $ ID                      : num  24 78 87 110 266 302 348 353 461 491 ...
 $ Age                     : num  21 45 23 22 29 18 47 48 26 38 ...
 $ Attrition               : chr  "No" "No" "No" "No" ...
 $ Early Attrition?        : chr  "No" "No" "No" "No" ...
 $ BusinessTravel          : chr  "Travel_Rarely" "Travel_Rarely" "Travel_Rarely" "Travel_Rarely" ...
 $ DailyRate               : num  391 193 541 534 1210 ...
 $ Department              : chr  "Research & Development" "Research & Development" "Sales" "Research & Develo
pment" ...
 $ DistanceFromHome        : num  15 6 2 15 2 10 4 29 29 1 ...
 $ Education               : num  2 4 1 3 3 3 1 1 2 1 ...
 $ EducationField          : chr  "Life Sciences" "Other" "Technical Degree" "Medical" ...
 $ EmployeeCount           : num  1 1 1 1 1 1 1 1 1 1 ...
 $ EmployeeNumber          : num  30 101 113 144 366 411 467 473 618 662 ...
 $ EnvironmentSatisfaction : num  3 4 3 2 1 4 2 1 1 3 ...
 $ Gender                  : chr  "Male" "Male" "Male" "Female" ...
 $ HourlyRate              : num  96 52 62 59 78 69 99 91 45 43 ...
 $ JobInvolvement          : num  3 3 3 3 2 2 3 3 3 3 ...
 $ JobLevel                : num  1 3 1 1 2 1 2 3 2 1 ...
 $ JobRole                 : chr  "Research Scientist" "Research Director" "Sales Representative" "Laboratory
Technician" ...
 $ JobSatisfaction         : num  4 1 1 4 2 3 3 3 3 1 ...
 $ MaritalStatus           : chr  "Single" "Married" "Divorced" "Single" ...
 $ MonthlyIncome           : num  1232 13245 2322 2871 6644 ...
 $ MonthlyRate             : num  19281 15067 9518 23785 3687 ...
 $ NumCompaniesWorked      : num  1 4 3 1 2 1 3 3 5 3 ...
 $ Over18                  : chr  "Y" "Y" "Y" "Y" ...
 $ OverTime                : chr  "No" "Yes" "No" "No" ...
 $ PercentSalaryHike       : num  14 14 13 15 19 12 19 21 12 17 ...
 $ PerformanceRating       : num  3 3 3 3 3 3 3 4 3 3 ...
 $ RelationshipSatisfaction: num  4 2 3 3 2 1 1 2 1 4 ...
 $ StandardHours           : num  80 80 80 80 80 80 80 80 80 80 ...
 $ StockOptionLevel        : num  0 0 1 0 2 0 0 1 2 0 ...
 $ TotalWorkingYears       : num  0 17 3 1 10 0 5 15 8 8 ...
 $ TrainingTimesLastYear   : num  6 3 3 5 2 2 3 3 5 3 ...
 $ WorkLifeBalance         : num  3 4 3 3 3 3 3 1 3 2 ...
```

```
$ YearsAtCompany          : num  0 0 0 0 0 0 0 0 0 0 ...
$ YearsInCurrentRole      : num  0 0 0 0 0 0 0 0 0 0 ...
$ YearsSinceLastPromotion : num  0 0 0 0 0 0 0 0 0 0 ...
$ YearsWithCurrManager    : num  0 0 0 0 0 0 0 0 0 0 ...
```

Data Selection

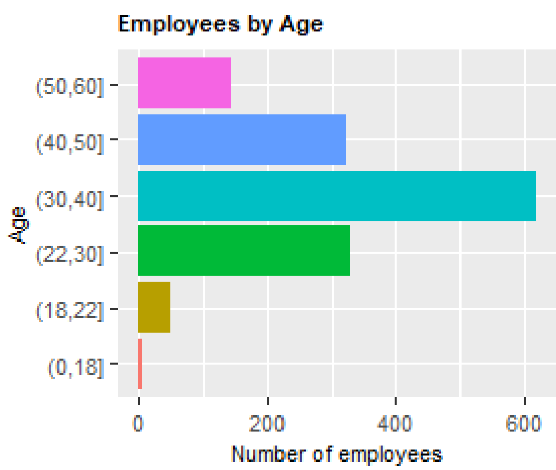
The following data elements were removed from the analysis:

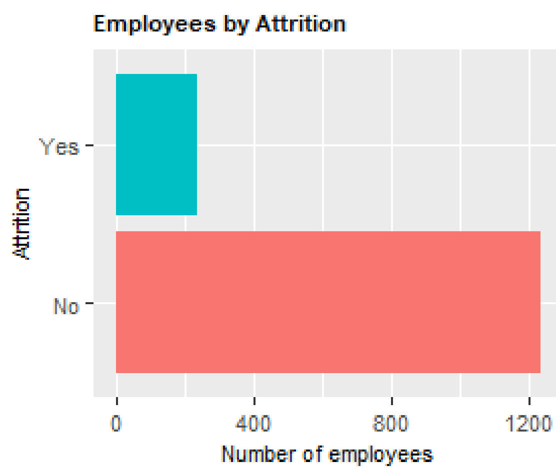
- DailyRate – Monthly income was used to represent employee earnings. This data element was redundant as a result.
- HourlyRate – Monthly income was used to represent employee earnings. This data element was redundant as a result.
- MonthlyRate – Monthly income was used to represent employee earnings. This data element was redundant as a result.
- Over18 – All employees are over 18, so this has no value. Age is a better variable.
- EmployeeCount – Each employee was given a value of “1” in this field.
- EmployeeNumber – This is the employee identification number and cannot add value to the analysis.
- StandardHours – All employees were given a value of 80.

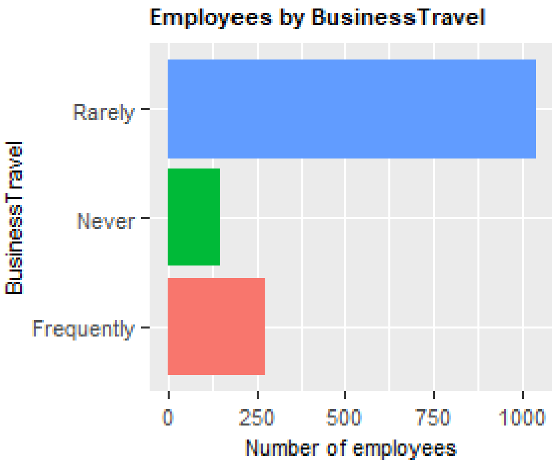
Some of the values could be considered convergent to represent a larger concept. For example, Relationship Satisfaction, Job Satisfaction and Job Involvement could separately or collectively act as a proxy for an “employee engagement” measure. Since there is not enough detail to understand how these measures were derived in the first place, this analysis will retain each measure separately and not make any assumptions regarding convergence.

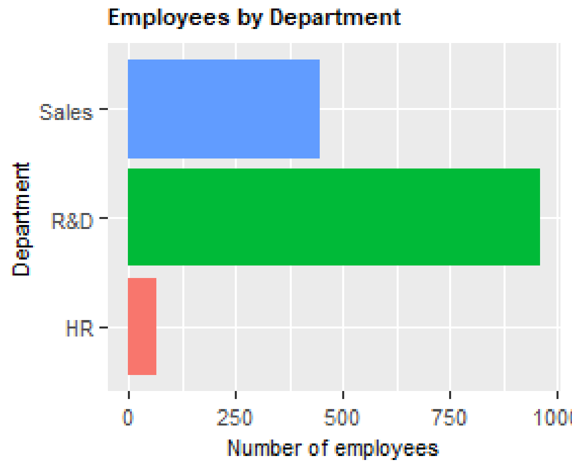
Data Visualizations

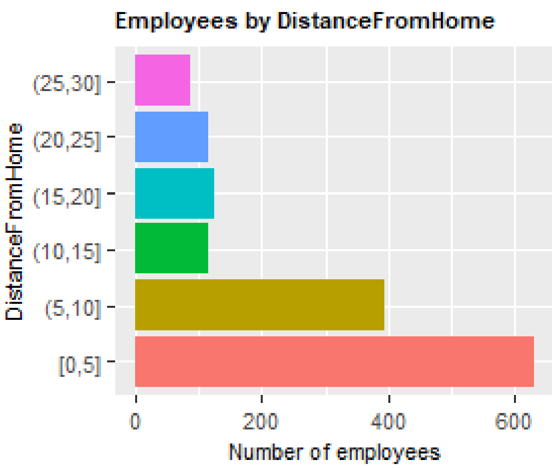
Each data element is visualized to provide visual context and build understanding of the data.

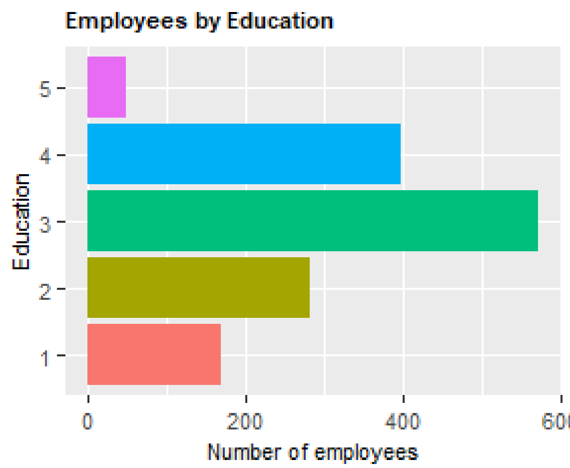


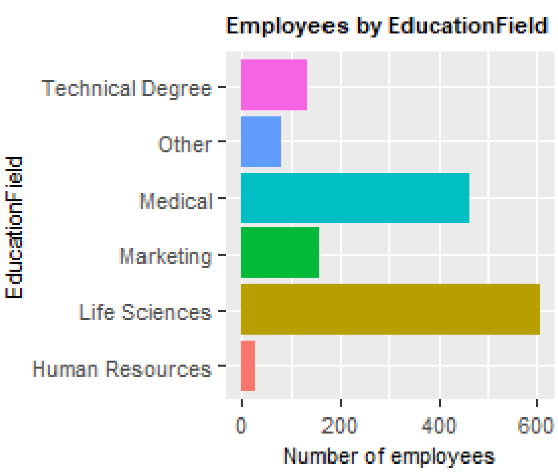


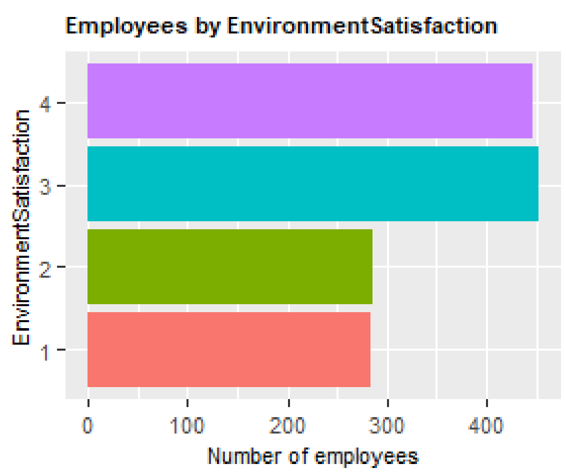

Employees by BusinessTravel
BusinessTravel
Rarely
Never
Frequently
0
250
500
750
1000
Number of employees

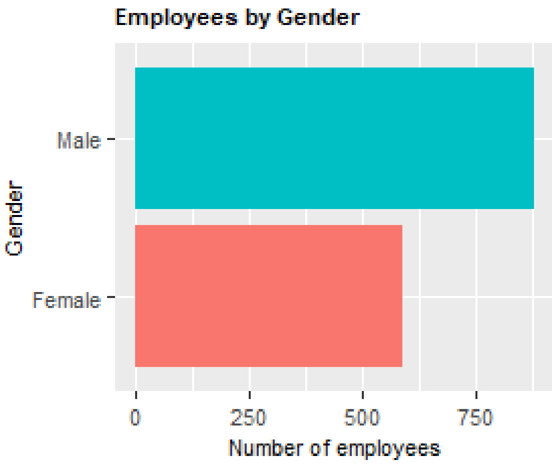

Employees by Department
Department
Sales
R&D
HR
0
250
500
750
Number of employees

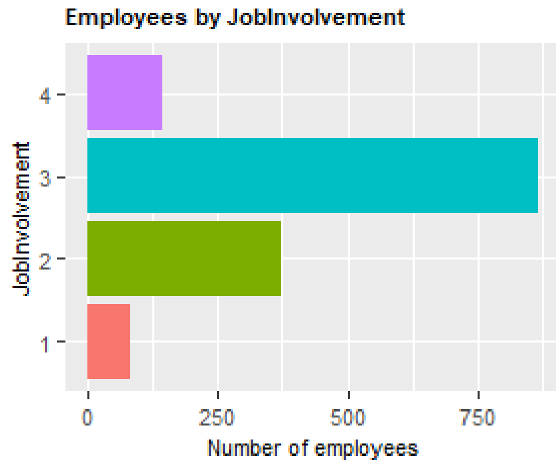

Employees by DistanceFromHome
DistanceFromHome
(25,30]
(20,25]
(15,20]
(10,15]
(5,10]
[0,5]
0
200
400
600
Number of employees

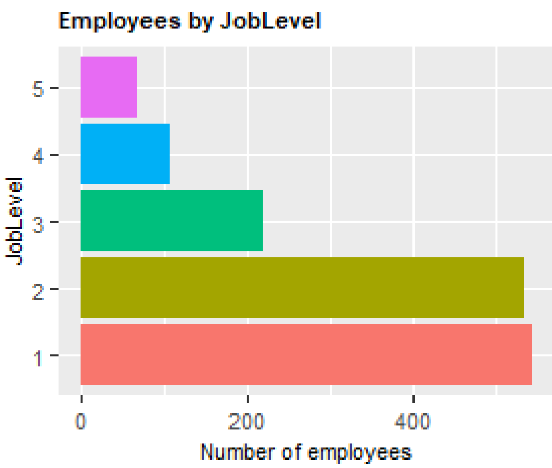

Employees by Education
Education
5
4
3
2
1
0
200
400
600
Number of employees

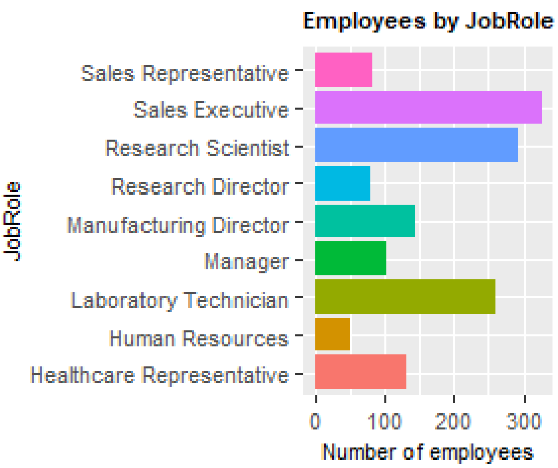

Employees by EducationField
EducationField
Technical Degree
Other
Medical
Marketing
Life Sciences
Human Resources
0
200
400
600
Number of employees

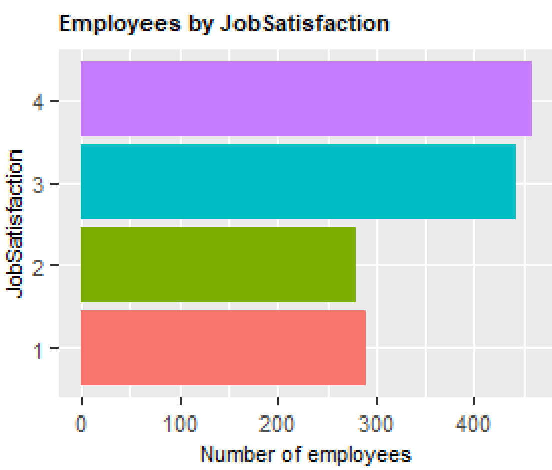

Employees by EnvironmentSatisfaction
EnvironmentSatisfaction
4
3
2
1
0
100
200
300
400
Number of employees

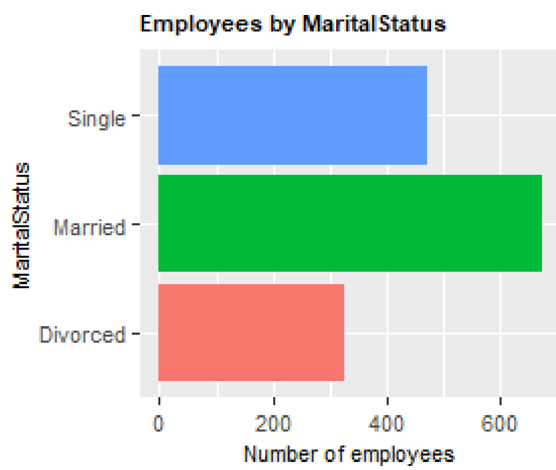

Employees by Gender
Gender
Male
Female
0
250
500
750
Number of employees

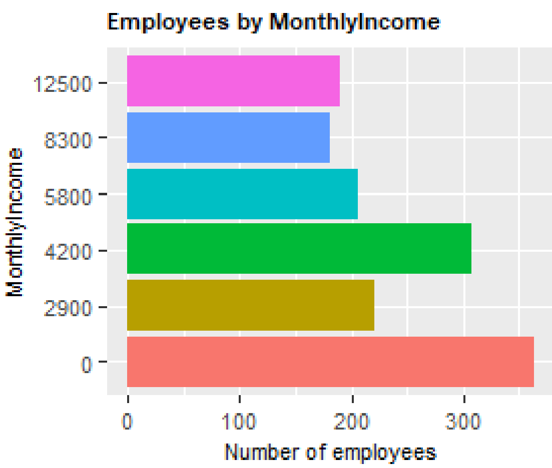

Employees by JobInvolvement
JobInvolvement
4
3
2
1
0
250
500
750
Number of employees

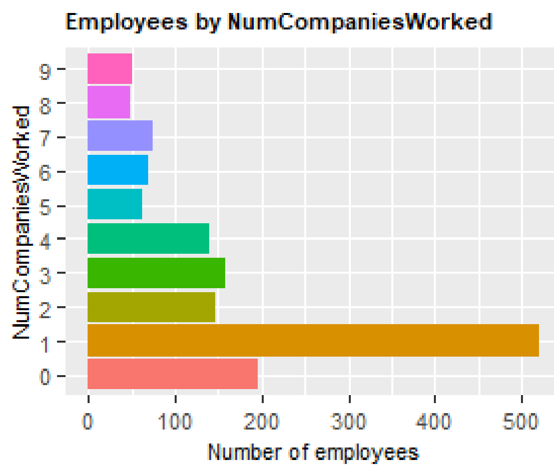

Employees by JobLevel
JobLevel
5
4
3
2
1
0
200
400
Number of employees

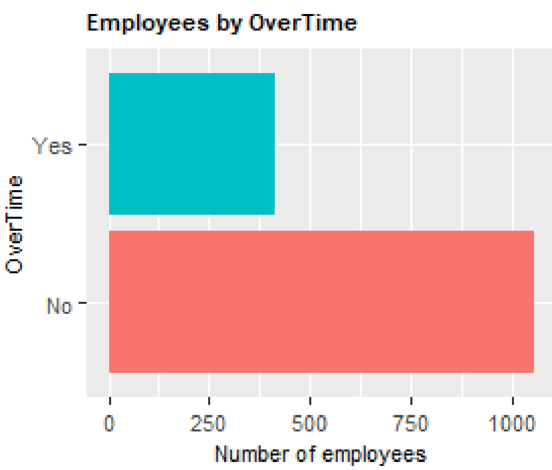

Employees by JobRole
JobRole
Sales Representative
Sales Executive
Research Scientist
Research Director
Manufacturing Director
Manager
Laboratory Technician
Human Resources
Healthcare Representative
0
100
200
300
Number of employees

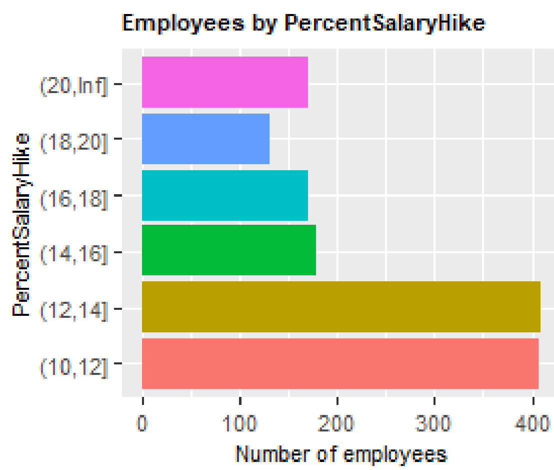

Employees by JobSatisfaction
JobSatisfaction
4
3
2
1
0
100
200
300
400
Number of employees

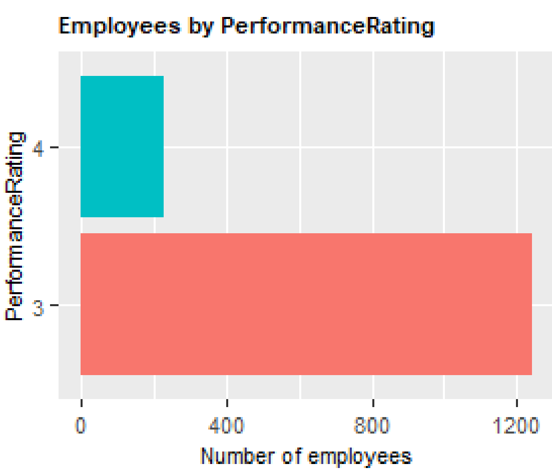

Employees by MaritalStatus
MaritalStatus
Single
Married
Divorced
0
200
400
600
Number of employees

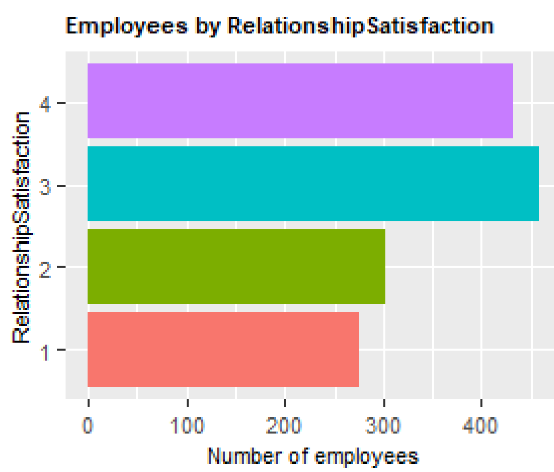

Employees by MonthlyIncome
MonthlyIncome
12500
8300
5800
4200
2900
0
0
100
200
300
Number of employees

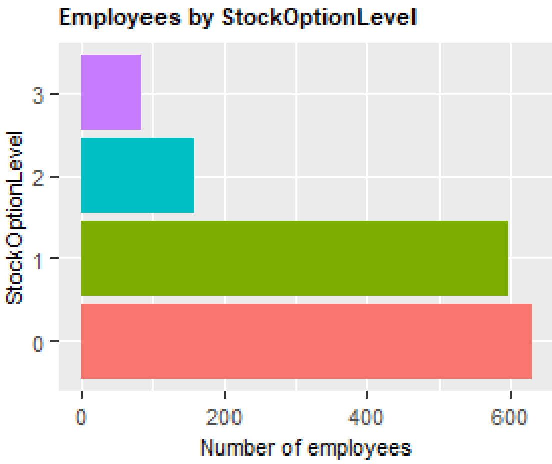

Employees by NumCompaniesWorked
NumCompaniesWorked
9
8
7
6
5
4
3
2
1
0
0
100
200
300
400
500
Number of employees

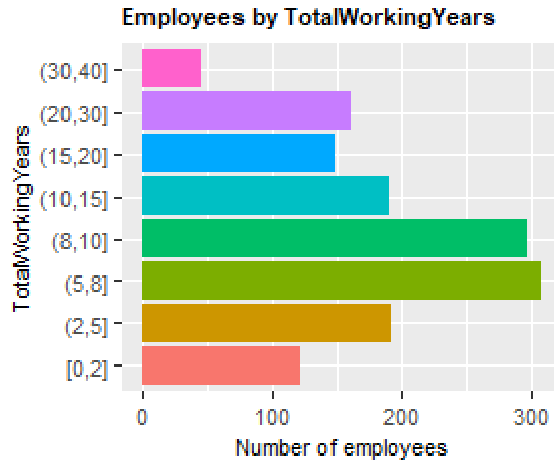

Employees by OverTime
OverTime
Yes
No
0
250
500
750
1000
Number of employees

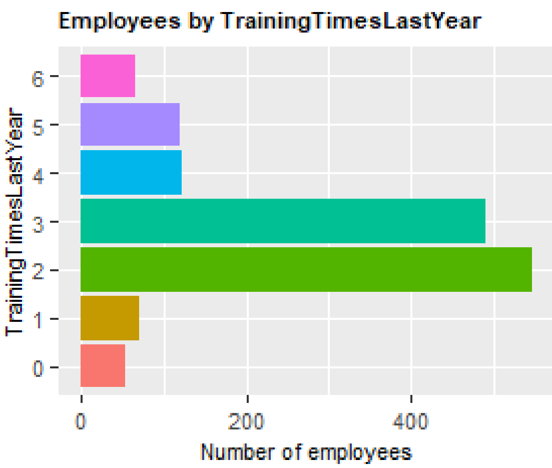

Employees by PercentSalaryHike
PercentSalaryHike
(20,Inf]
(18,20]
(16,18]
(14,16]
(12,14]
(10,12]
0
100
200
300
400
Number of employees

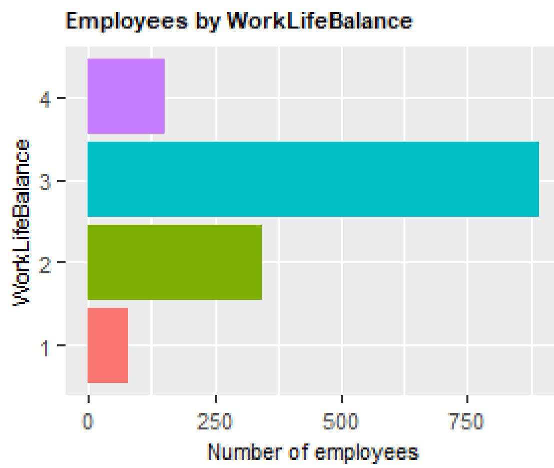

Employees by PerformanceRating
PerformanceRating
4
3
0
400
800
1200
Number of employees

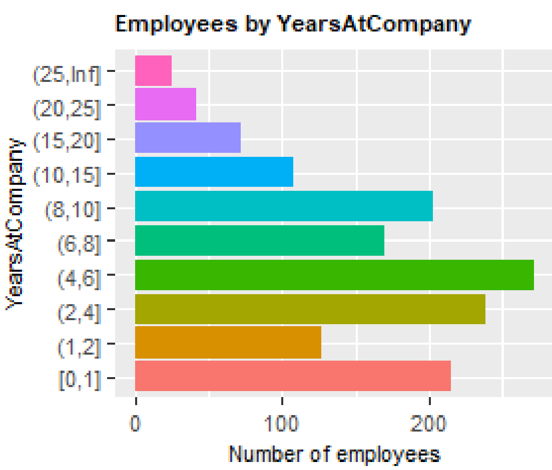

Employees by RelationshipSatisfaction
RelationshipSatisfaction
4
3
2
1
0
100
200
300
400
Number of employees

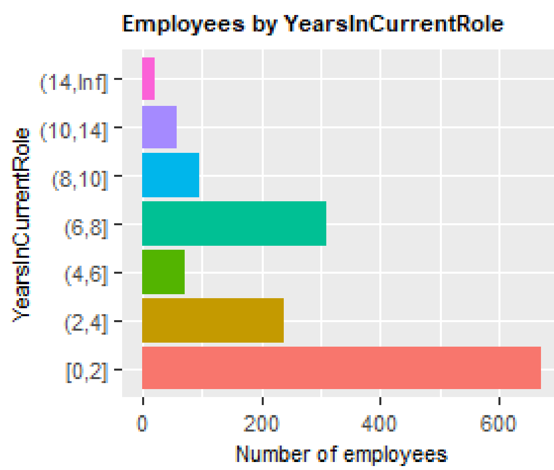

Employees by StockOptionLevel
StockOptionLevel
3
2
1
0
0
200
400
600
Number of employees

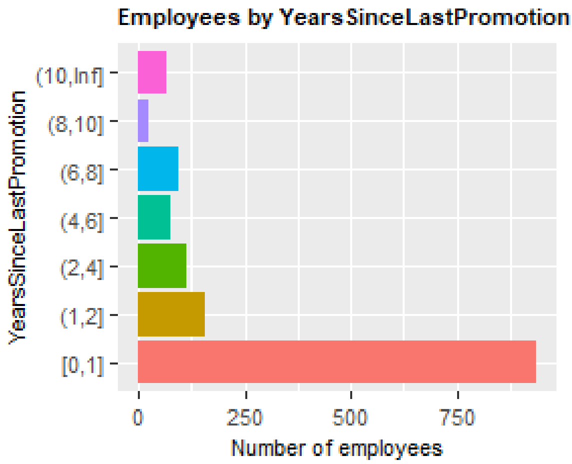

Employees by TotalWorkingYears
TotalWorkingYears
(30,40]
(20,30]
(15,20]
(10,15]
(8,10]
(5,8]
(2,5]
[0,2]
0
100
200
300
Number of employees

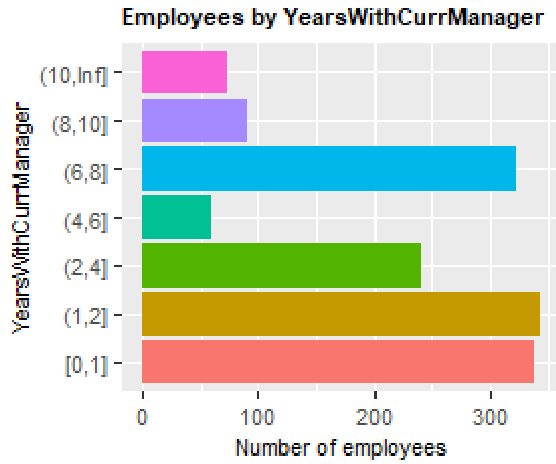

Employees by TrainingTimesLastYear
TrainingTimesLastYear
6
5
4
3
2
1
0
0
200
400
Number of employees


Employees by WorkLifeBalance
WorkLifeBalance
4
3
2
1
0
250
500
750
Number of employees


Employees by YearsAtCompany
YearsAtCompany
(25,Inf]
(20,25]
(15,20]
(10,15]
(8,10]
(6,8]
(4,6]
(2,4]
(1,2]
[0,1]
0
100
200
Number of employees


Employees by YearsInCurrentRole
YearsInCurrentRole
(14,Inf]
(10,14]
(8,10]
(6,8]
(4,6]
(2,4]
[0,2]
0
200
400
600
Number of employees





Data Transformation

After initial review of the data, the following variables were discretized:

Age – 0, 18, 22, 30, 40, 50, 60, Inf. The logic was that there are no employees under 18, the period from 18-22 can be considered college years, 22-30 is post college/twenties, then by decade.

DistanceFromHome – Since this is a continuous variable, we discretized to 0, 5, 10, 15, 20, 25, 30, Inf. To group by commuting distance.

MonthlyIncome – This was a continuous variable that was discretized to 0, 2900, 4200, 5800, 8300, 12500, Inf to roughly corelate to annual income buckets of $0, $35000, $50000, $70000, $100000, $150000, Inf

PercentSalaryHike – Discretized to 0, 10, 12, 14, 16, 18, 20, Inf. While these may seem like high annual increase percentages, the breakout was consistent with the data which was also high.

TotalWorkingYears – Discretized to 0, 2, 5, 8, 10, 15, 20, 30, 40, Inf to allow for analysis of turnover in the early years of employment.

YearsAtCompany – Discretized to 0, 1, 2, 4, 6, 8, 10, 15, 20, 25, Inf

YearsInCurrentRole – Discretized to 0, 2, 4, 6, 8, 10, 14, Inf

YearsSinceLastPromotion - Discretized to 0, 1, 2, 4, 6, 8, 10, Inf

YearsWithCurrManage – Discretized to 0, 1, 2, 4, 6, 8, 10, Inf

Additional Transformation

After reviewing the records for attrition, a clear pattern emerge that shows that attrition is much more prevalent in the early years of employment (Figure 3). As discussed in the introduction, attrition earlier in the employee lifecycle can be disproportionately expensive, the decision was made to investigate early attrition as part of this analysis. To facilitate the analysis, additional variables were added to the dataset to denote whether the employee left the company before completing 2 years and 5 years of service respectively.

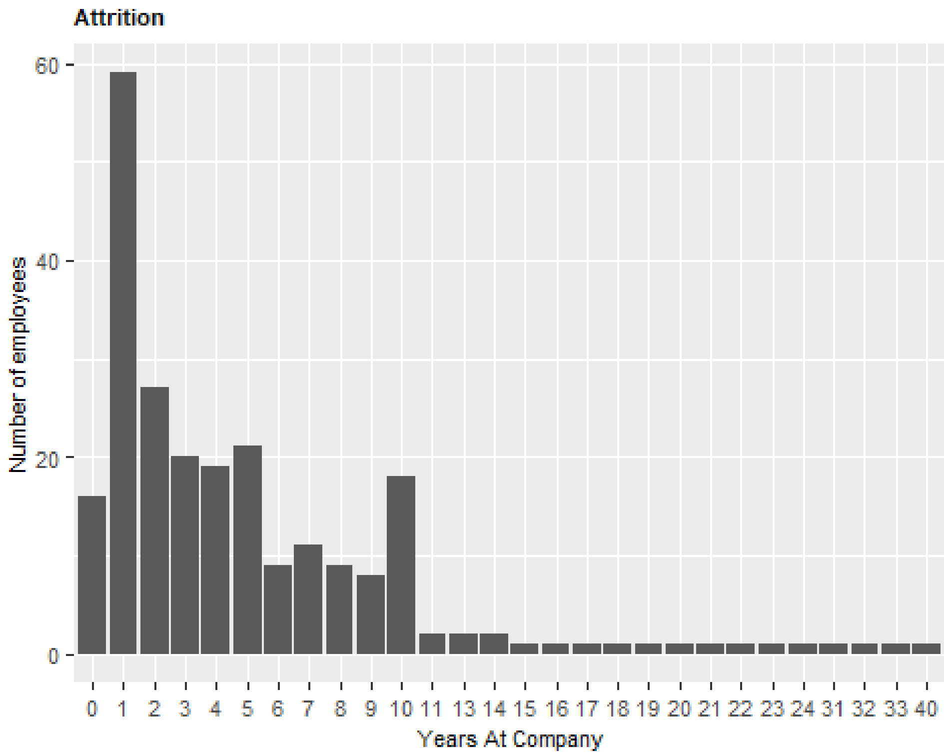


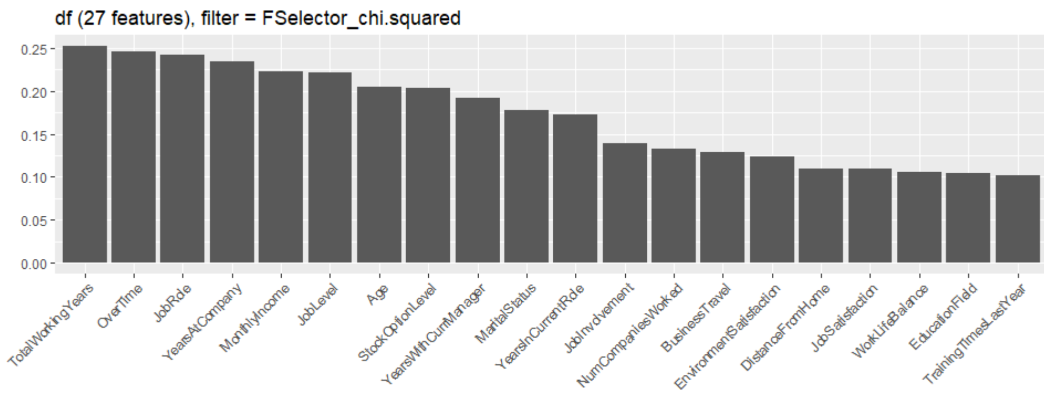
Figure 3 – Employee Attrition by years of service

Feature Selection

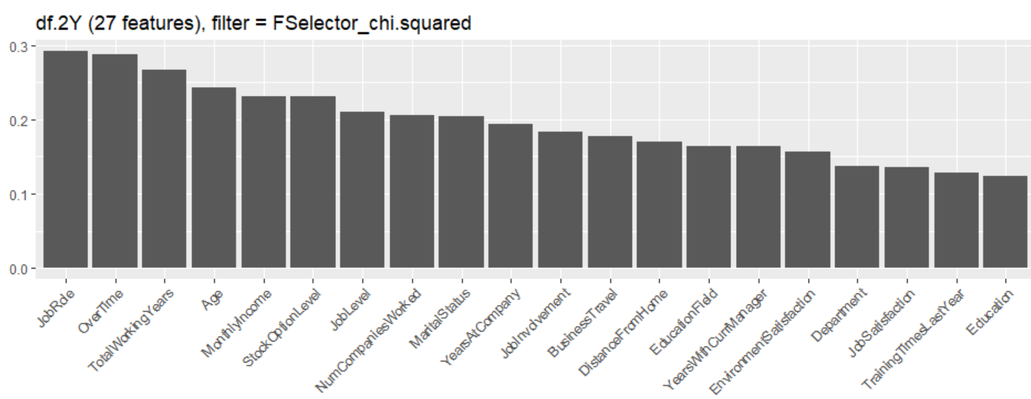
Using Chi squared, the variables with the highest correlation to attrition were identified. All 27 variables were used for filtering. The top 20 variables for each of the three employee tenure populations are displayed. While feature selection is not used directly in the learning algorithms, this information is helpful to understand what variables are important relative to one another.

Note that total working years has the highest correlation when looking at the total workforce. This reinforced the value of tenure and the desire to break out the less than 2 year and less than 5 year turnover populations.

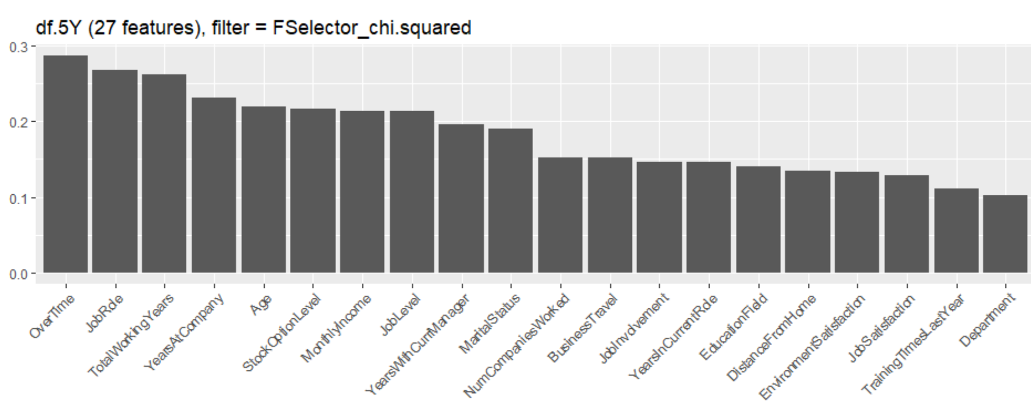
Feature selection for the entire employee population:



Feature selection for employees leaving with less than 2 years of service:



Feature selection for employees leaving with less than 5 years of service:



3. Unsupervised modeling

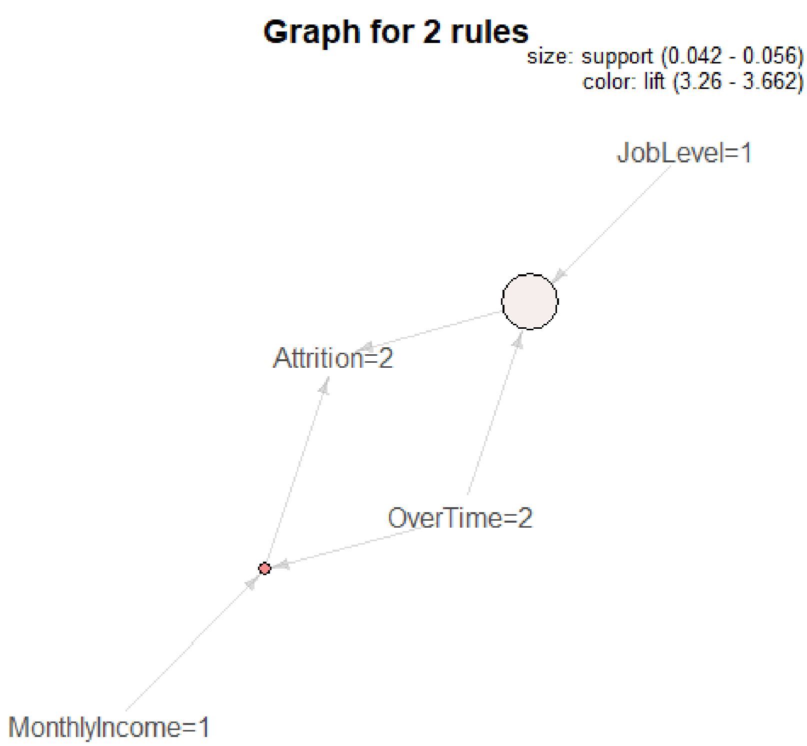
Unsupervised learning is used as another way to look at the data and obtain an idea of the relative importance between variables and to support the features selected. For example, in the Association Rules Mining output, the LHS variables are the ones that appear as the highest on the feature selection graphs.

<u>Association Rule Mining</u>

The Apriori algorithm was used to perform association rules mining on the data. Positive attrition was placed on the right hand and with support >= 0.03, and confidence >= 0.5. These values were selected after experimenting with different ranges and provided the best set of results across all three employee populations.

Association Rules – Entire Population

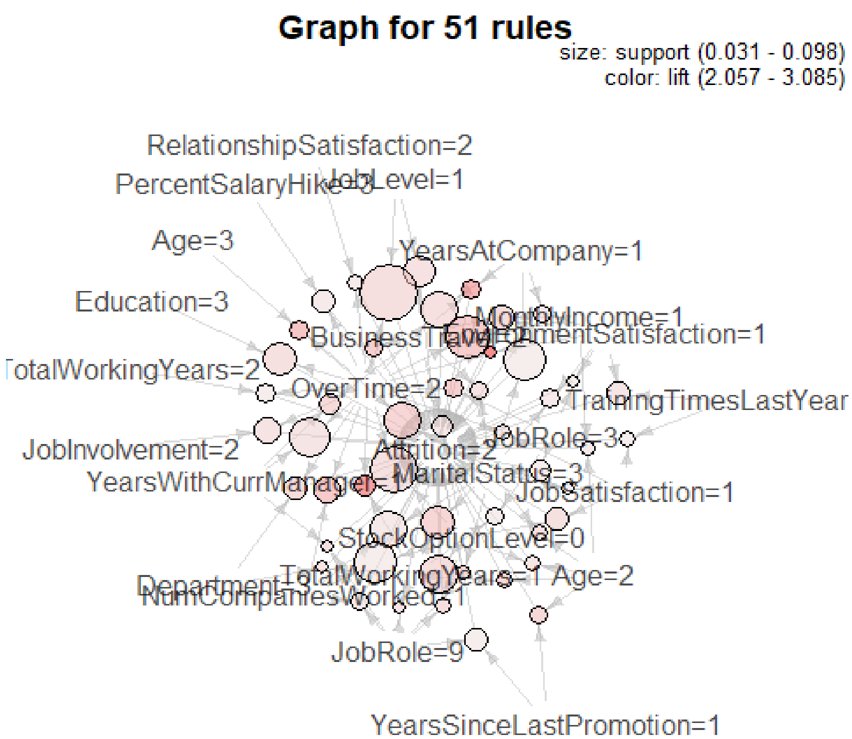
lhs	Rhs	support	confidence	lift	count
{MonthlyIncome=1,OverTime=2}	{Attrition=2}	0.042	0.590	3.662	62
{JobLevel=1,OverTime=2}	{Attrition=2}	0.056	0.526	3.260	82



Association Rules – Employees leaving with less than 2 years of service:

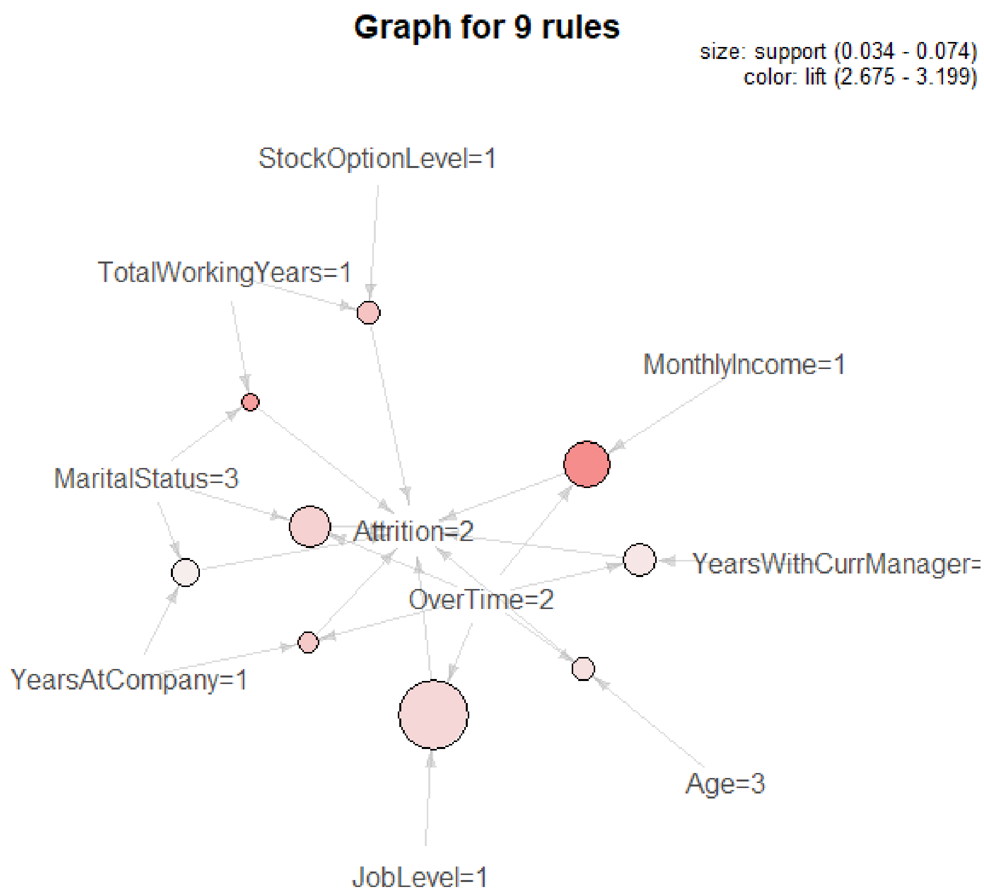
Lhs	rhs	support	confidence	lift	count
{EnvironmentSatisfaction=1,OverTime=2}	{Attrition=2}	0.031	0.750	3.085	18
{OverTime=2,TotalWorkingYears=1}	{Attrition=2}	0.043	0.714	2.938	25
{BusinessTravel=2,YearsAtCompany=1}	{Attrition=2}	0.041	0.667	2.742	24
{Age=3,OverTime=2}	{Attrition=2}	0.040	0.639	2.628	23
{NumCompaniesWorked=1,OverTime=2}	{Attrition=2}	0.052	0.625	2.571	30
{MonthlyIncome=1,OverTime=2}	{Attrition=2}	0.078	0.616	2.536	45
{BusinessTravel=2,OverTime=2}	{Attrition=2}	0.038	0.611	2.514	22
{JobRole=3,OverTime=2}	{Attrition=2}	0.040	0.605	2.490	23
{MaritalStatus=3,OverTime=2}	{Attrition=2}	0.066	0.585	2.405	38
{MaritalStatus=3,TotalWorkingYears=1}	{Attrition=2}	0.060	0.583	2.400	35
{Age=2,YearsSinceLastPromotion=1}	{Attrition=2}	0.038	0.579	2.381	22
{Age=2,TotalWorkingYears=1}	{Attrition=2}	0.034	0.571	2.351	20
{BusinessTravel=2,YearsWithCurrManager=1}	{Attrition=2}	0.043	0.568	2.337	25
{Department=3,OverTime=2}	{Attrition=2}	0.047	0.562	2.314	27
{BusinessTravel=2,MonthlyIncome=1}	{Attrition=2}	0.048	0.560	2.304	28
{JobRole=9,MaritalStatus=3}	{Attrition=2}	0.033	0.559	2.299	19
{StockOptionLevel=0,TotalWorkingYears=1}	{Attrition=2}	0.069	0.556	2.285	40
{JobLevel=1,OverTime=2}	{Attrition=2}	0.098	0.553	2.276	57
{JobRole=9,StockOptionLevel=0}	{Attrition=2}	0.036	0.553	2.273	21
{OverTime=2,YearsAtCompany=1}	{Attrition=2}	0.066	0.551	2.265	38
{JobRole=9,YearsWithCurrManager=1}	{Attrition=2}	0.031	0.545	2.244	18
{Education=3,OverTime=2}	{Attrition=2}	0.062	0.545	2.244	36
{OverTime=2,StockOptionLevel=0}	{Attrition=2}	0.083	0.545	2.244	48
{JobSatisfaction=1,StockOptionLevel=0}	{Attrition=2}	0.045	0.542	2.228	26
{Age=2,MaritalStatus=3}	{Attrition=2}	0.034	0.541	2.224	20
{BusinessTravel=2,JobLevel=1}	{Attrition=2}	0.059	0.540	2.220	34
{JobInvolvement=2,OverTime=2}	{Attrition=2}	0.050	0.537	2.209	29
{BusinessTravel=2,MaritalStatus=3}	{Attrition=2}	0.038	0.537	2.207	22
{OverTime=2,YearsWithCurrManager=1}	{Attrition=2}	0.072	0.532	2.187	42
{JobRole=9,TotalWorkingYears=1}	{Attrition=2}	0.031	0.529	2.178	18
{Department=3,TotalWorkingYears=1}	{Attrition=2}	0.031	0.529	2.178	18
{EnvironmentSatisfaction=1,TrainingTimesLastYear=2}	{Attrition=2}	0.047	0.529	2.178	27
{Age=2,StockOptionLevel=0}	{Attrition=2}	0.036	0.525	2.160	21
{EnvironmentSatisfaction=1,YearsAtCompany=1}	{Attrition=2}	0.038	0.524	2.155	22
{BusinessTravel=2,StockOptionLevel=0}	{Attrition=2}	0.043	0.521	2.142	25
{Age=2,MonthlyIncome=1}	{Attrition=2}	0.033	0.514	2.112	19
{MaritalStatus=3,NumCompaniesWorked=1}	{Attrition=2}	0.067	0.513	2.111	39
{OverTime=2,RelationshipSatisfaction=2}	{Attrition=2}	0.034	0.513	2.109	20

{JobSatisfaction=1,OverTime=2}	{Attrition=2}	0.034	0.513	2.109	20
{JobRole=9,NumCompaniesWorked=1}	{Attrition=2}	0.038	0.512	2.105	22
{EnvironmentSatisfaction=1,MaritalStatus=3}	{Attrition=2}	0.040	0.511	2.102	23
{OverTime=2,PercentSalaryHike=3}	{Attrition=2}	0.045	0.510	2.097	26
{JobRole=9,YearsSinceLastPromotion=1}	{Attrition=2}	0.047	0.500	2.057	27
{JobSatisfaction=1,MaritalStatus=3}	{Attrition=2}	0.031	0.500	2.057	18
{JobSatisfaction=1,TrainingTimesLastYear=2}	{Attrition=2}	0.034	0.500	2.057	20
{JobRole=3,TotalWorkingYears=1}	{Attrition=2}	0.038	0.500	2.057	22
{EnvironmentSatisfaction=1,JobRole=3}	{Attrition=2}	0.031	0.500	2.057	18
{JobRole=3,MaritalStatus=3}	{Attrition=2}	0.043	0.500	2.057	25
{OverTime=2,TotalWorkingYears=2}	{Attrition=2}	0.040	0.500	2.057	23
{NumCompaniesWorked=1,StockOptionLevel=0}	{Attrition=2}	0.076	0.500	2.057	44
{MaritalStatus=3,YearsAtCompany=1}	{Attrition=2}	0.076	0.500	2.057	44



Association Rules – Employees leaving with less than 5 years of service:

lhs	rhs	support	confidence	lift	count
{MonthlyIncome=1,OverTime=2}	{Attrition=2}	0.057	0.598	3.199	58
{MaritalStatus=3,TotalWorkingYears=1}	{Attrition=2}	0.034	0.583	3.121	35
{StockOptionLevel=1,TotalWorkingYears=1}	{Attrition=2}	0.039	0.556	2.973	40
{OverTime=2,YearsAtCompany=1}	{Attrition=2}	0.037	0.551	2.947	38
{MaritalStatus=3,OverTime=2}	{Attrition=2}	0.053	0.540	2.889	54
{JobLevel=1,OverTime=2}	{Attrition=2}	0.074	0.531	2.844	76
{Age=3,OverTime=2}	{Attrition=2}	0.038	0.520	2.782	39
{OverTime=2,YearsWithCurrManager=1}	{Attrition=2}	0.047	0.511	2.732	48
{MaritalStatus=3,YearsAtCompany=1}	{Attrition=2}	0.043	0.500	2.675	44



ARM Observations

It is interesting that the population of employees leaving with less than 2 years of service had 51 rules vs. 2 rules for the entire population. One should not infer that there are 25 times more combinations of drivers for this population. Instead, one should remember that with a smaller population sample to analyze, all variable become more important. Remember that the calculation for ARM for the smaller population will have a smaller denominator as a result, thus driving up the values.

Clustering

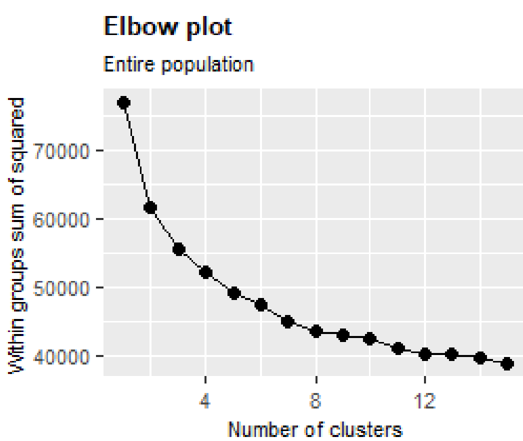
Clustering is another way to visualize and understand the relationships within the data set. The k-means algorithm is sensitive to the randomly-chosen cluster centers, so selecting a k-value to use to seed the clustering analysis is critical. Setting k too high will improve the homogeny of the clusters, but it risks overfitting the data. Setting it too low has the opposite effect. Two methods are used in this analysis to demonstrate different approaches.

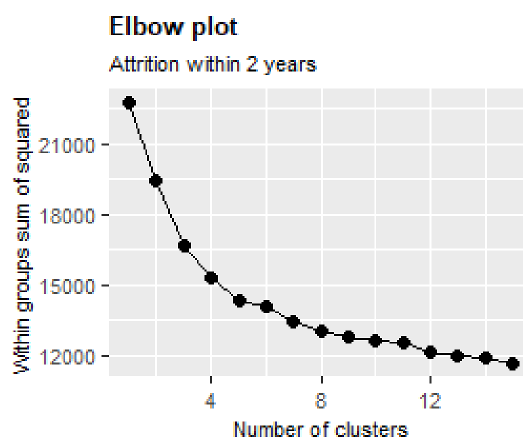
Elbow Method

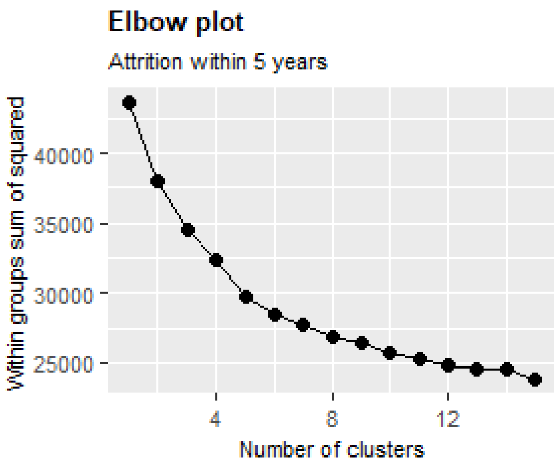
The elbow method can be used when there is no prior knowledge about the data. This method attempts to gauge how the homogeneity or heterogeneity within the clusters changes for various values of k. In a dataset, homogeneity within clusters is expected to increase as additional clusters are added. Conversely, heterogeneity will decrease with more clusters. By using R to statistically measure homogeneity and heterogeneity and plotting those results, a plot can be produces that allows one to find k so that there are diminishing returns beyond that point on the curve. That point is called the elbow.

The following elbows were created for each of the three employee populations:

Employee attrition – Entire population

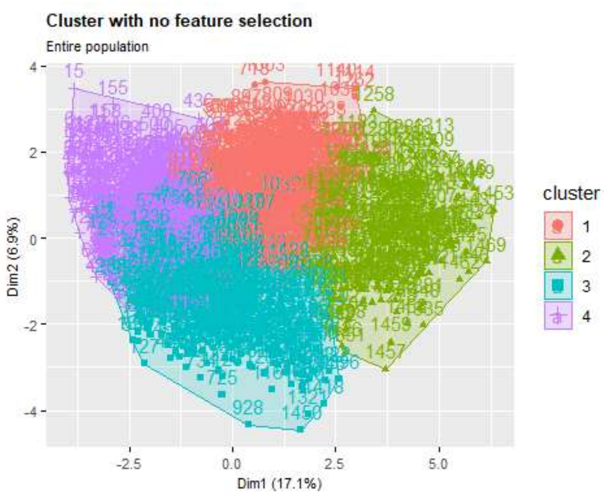


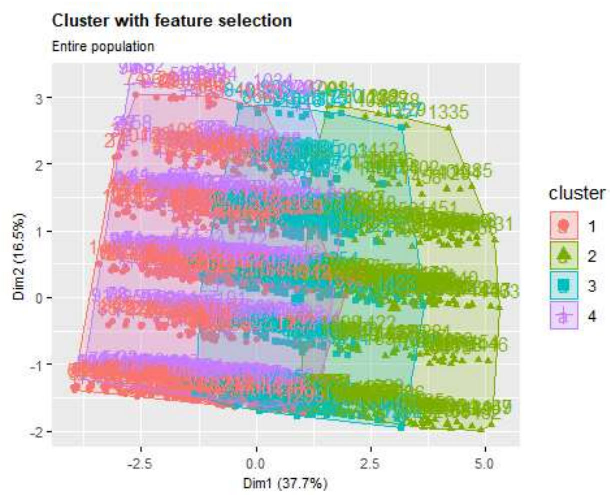




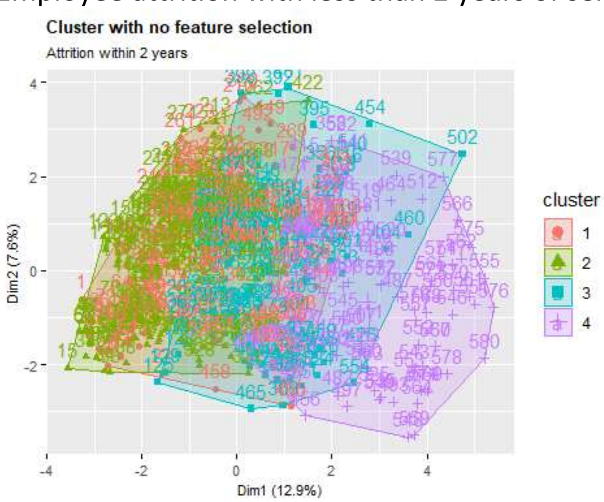
Since the elbow for the entire employee population is k=4, 4 was used for the following graphs. For no feature selection, all 27 features were included. For graphs with feature selection, only the top 10 features identified in the feature selection section were used:

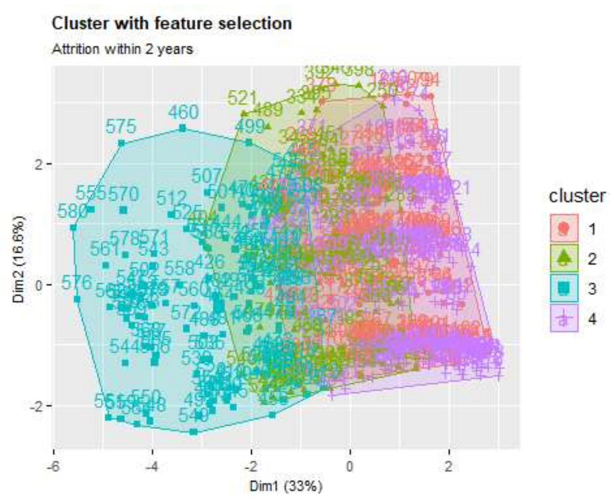
Employee attrition – Entire population



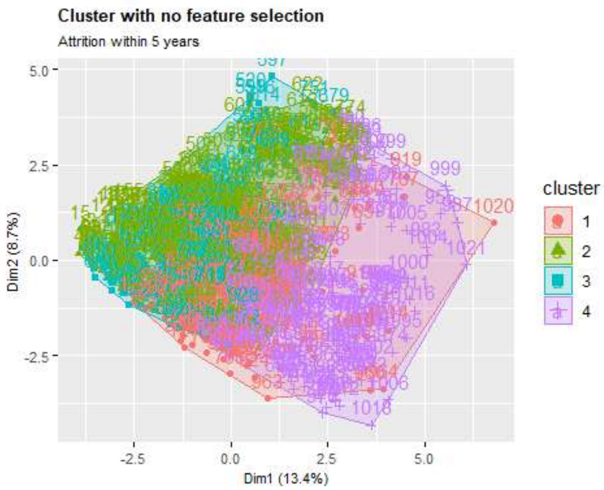


Employee attrition with less than 2 years of service





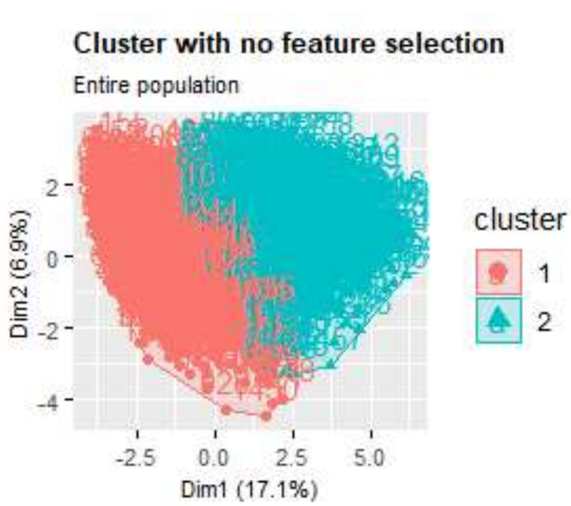
Employee attrition with less than 5 years of service

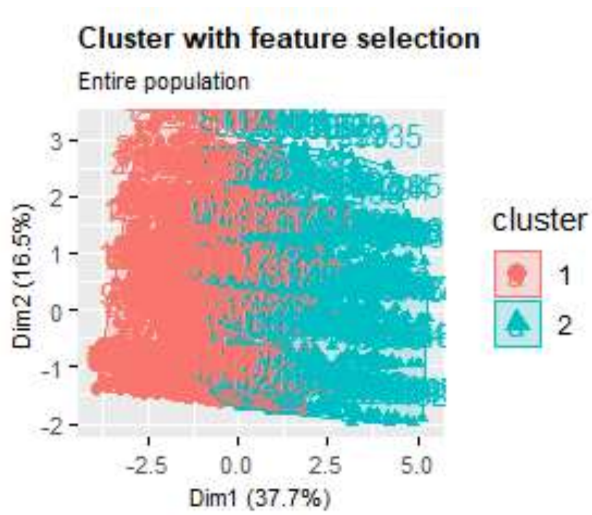


Prior Knowledge method

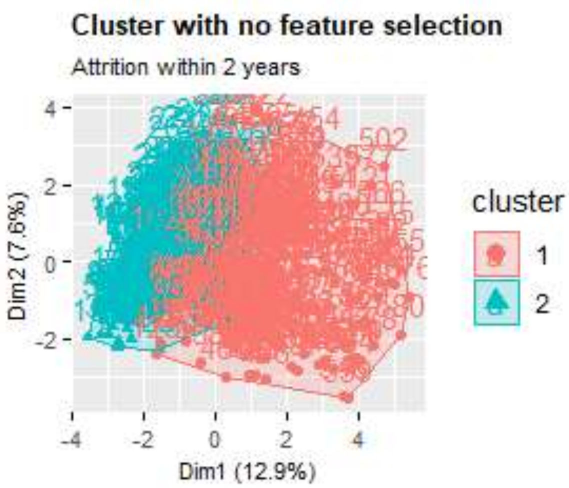
The second method is simple and requires that prior knowledge (a priori) about the data guide in the selection of k. In this dataset, the goal is to identify the variables that contribute to an employee leaving. Therefore, we are searching to understand two clusters that represent employees who leave and those who do not. In this case k=2.

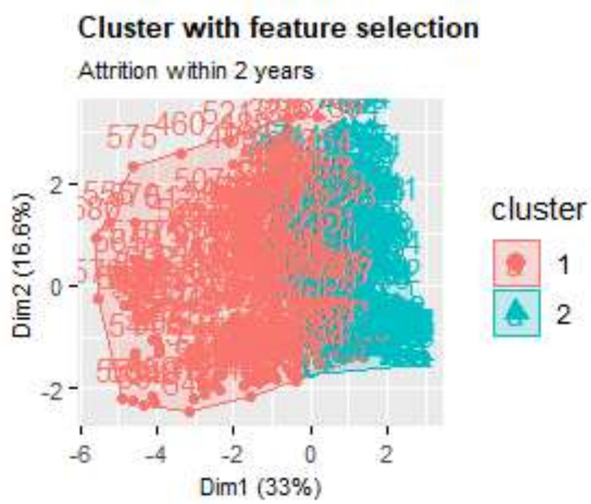
Employee attrition – Entire population



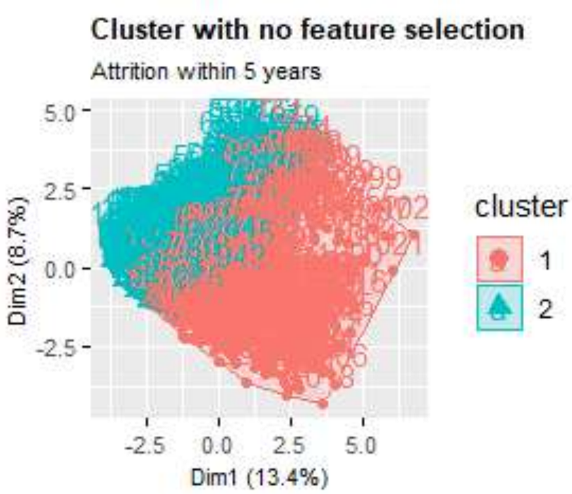


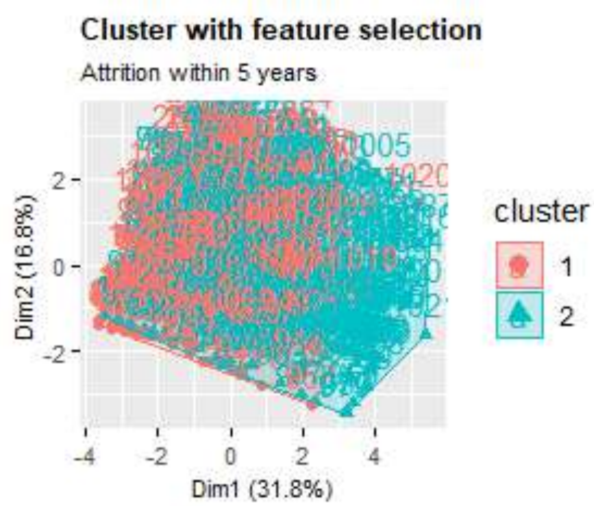
Employee attrition with less than 2 years of service





Employee attrition with less than 5 years of service





4. Models

The following 5 machine learning models were used to analyze the data:

- Naïve Bayes
- Decision Tree
- kNN
- Random Forest
- SVM

The model selection process was completed through nested resampling. This was done to ensure that the model is as unbiased as possible. This approach also completed feature selection as part of the model selection. Using the nested approach, the computer randomly chooses parameter values from an allowable range of data that is provided by the user.

Tuning

Hyperparameter tuning is the process of choosing a set of optimal hyperparameters for a learning algorithm. An actual hyperparameter is parameter whose value is set before the learning process begins. The approach used here is the nested approach introduced above:

Parameter tuning (the hyperparameters) and feature selection is accomplished within the inner loop and the performance is estimated with the outer loop. The tuning strategy is a random search with 100 iterations. The feature selection was done as part of the tuning, allowing for 3 to 10 predictors being used for each model. This approach then takes the best features and parameters and creates an "optimized model". This is completed for 100 variations of aforementioned variables. After the best model is determined, 5-fold cross-validation is performed to validate that the optimized model has the best performance.

The hyperparameters and the ranges for each parameter used for training each model are as follows:

Model	Parameter Range
Naïve Bayes	laplace: 0 to 5
Decision Tree	Complexity parameter: 10^(-8) to 1
kNN	2 to 5
Random Forest	Number of trees: 1 to 500
SVM (kernel: linear, polynomial, rbf)	Linear - Cost: 2^(-12) to 2^12 Polynomial - Cost: 2^(-12) to 2^12, Degree: 2 to 5 Rbf- Cost: 2^(-12) to 2^12, Sigma: 2^(-12) to 2^12

5. Results

Approach to performance

AUC - ROC curve is a performance measurement for classification problems at various thresholds settings. ROC (Receiver Operating Characteristics) is a probability curve and AUC (Area Under The Curve) represents degree or measure of separability. It tells how much model can distinguish between classes. The higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s. That means the higher the AUC, the better the model is at distinguishing between employees that terminate and those that do not.

Interpreting the results

An excellent model has an AUC value very close to 1. This denotes a good measure of separability. A model with an AUC near 0 is considered a poor model because it has a much worse measure of separability. Not only that, a low AUC means that the model is producing false positive and false negative results. An AUC value of 0.5, means model has no class separation capacity whatsoever. In other words, it tells you the very least.

Model results

Entire Population:

Model	Optimized parameters	auc	mmce	acc
Naive Bayes	# Predictors: 4 laplace: 0	77%	19%	81%
Decision Tree	# Predictors: 9 cp: 9.979 x 10^(-8)	80%	13%	87%
kNN	# Predictors: 5 k: 4	91%	10%	90%
Random Forest	*# Predictors: 6* *ntree: 452*	*96%*	*6%*	*94%*
SVM	# Predictors: 9 kernel: rbf C: 5.894 x 10^(-4)	79%	15%	85%

Less than 2 years of service:

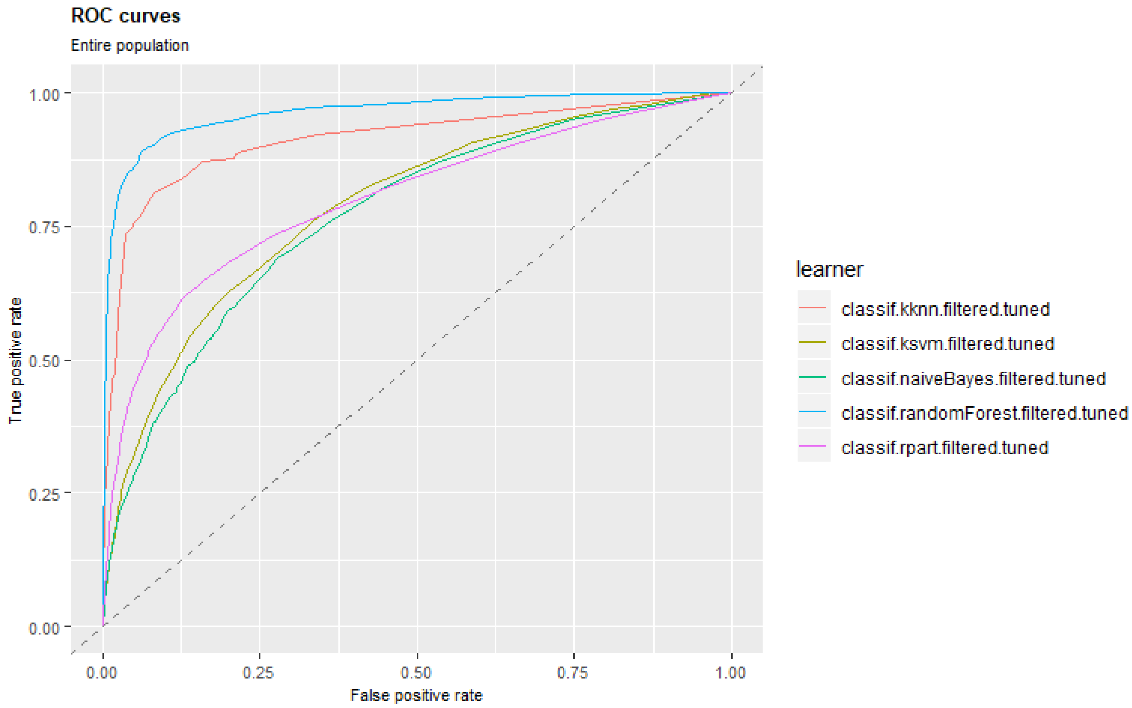
Model	Optimized parameters	auc	mmce	acc
Naive Bayes	# Predictors: 5 laplace: 3	79%	22%	78%
Decision Tree	# Predictors: 6 cp: 1.576 x 10^(-5)	80%	17%	83%
kNN	# Predictors: 10 k: 2	87%	14%	86%
Random Forest	*# Predictors: 7* *ntree: 482*	*91%*	*11%*	*89%*
SVM	# Predictors: 3 kernel: Linear C: 3.610	83%	19%	81%

Less than 5 years of service:

Model	Optimized parameters	auc	mmce	acc
Naive Bayes	# Predictors: 5 laplace: 0	80%	18%	82%
Decision Tree	# Predictors: 5 cp: 0.391	79%	14%	86%
kNN	*# Predictors: 10* *k: 3*	*94%*	*9%*	*91%*
Random Forest	# Predictors: 3 ntree: 474	94%	10%	90%
SVM	# Predictors: 8 kernel: Linear C: 0.159	84%	15%	85%

ROC Curves

ROC Curve – Employee turnover entire population

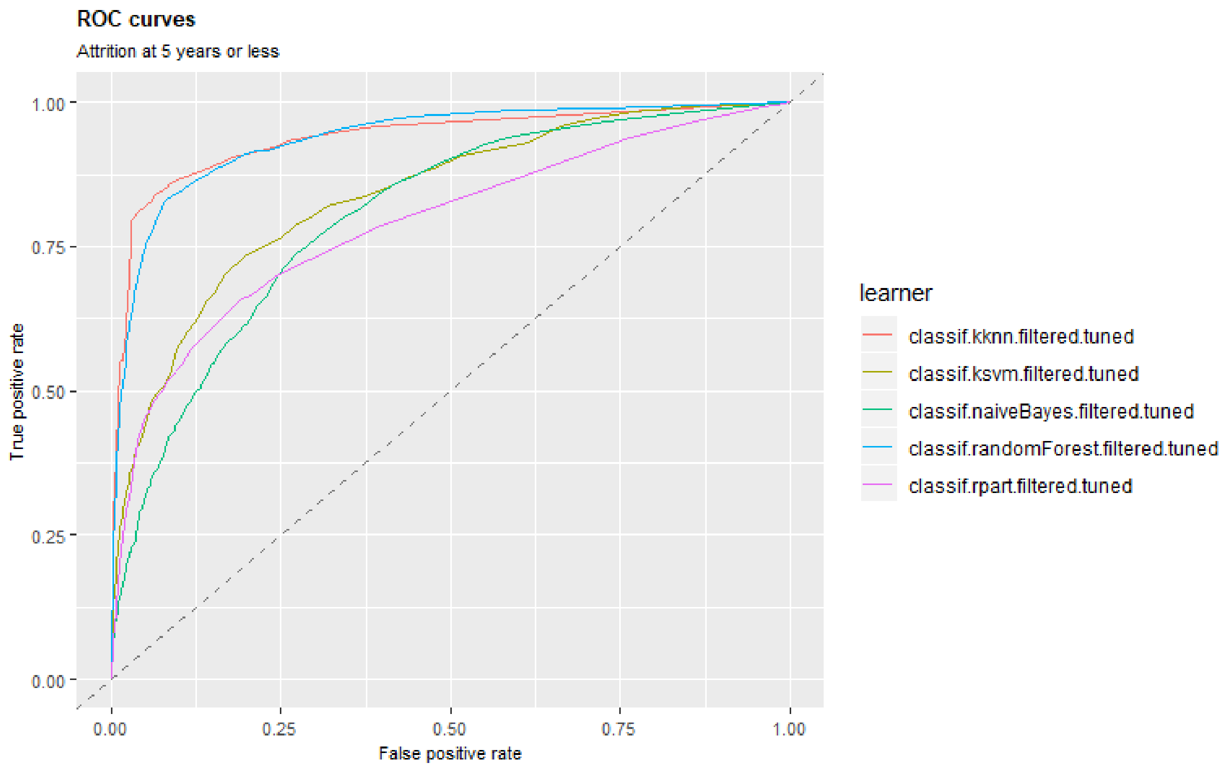



```
NB - Entire Pop
    predicted
true    1   2
   1 5441 724
   2  659 526
      auc      mmce       acc
0.7724678 0.1881633 0.8118367

 RPart - Entire Pop
    predicted
true    1   2
   1 5916 249
   2  693 492
      auc      mmce       acc
0.8008505 0.1281633 0.8718367

 kNN - Entire Pop
    predicted
true    1   2
   1 6037 128
   2  588 597
       auc       mmce        acc
0.91381906 0.09741497 0.90258503
```

```
RF - Entire Pop
    predicted
true    1   2
   1 6086  79
   2  333 852
       auc       mmce        acc
0.96401833 0.05605442 0.94394558

 SVM - Entire Pop
    predicted
true    1   2
   1 6140 25
   2 1105 80
      auc      mmce       acc
0.7886910 0.1537415 0.8462585
```

Observation
Random Forest produced the best result with an AUC of 96.40%.

ROC Curve – Employee turnover < 5 years



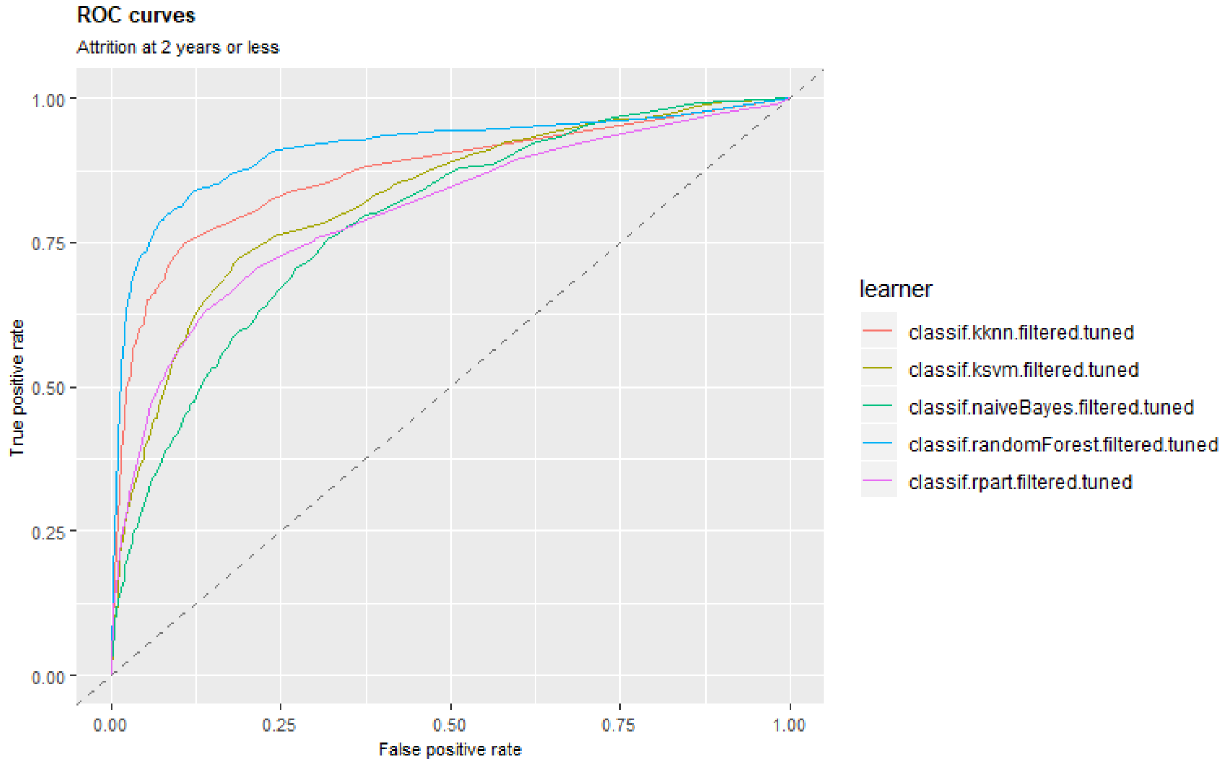

```
NB - 5 Year
    predicted
true     1    2
   1 3788 367
   2  550 405
        auc      mmce       acc
0.8042603 0.1794521 0.8205479

 RPart - 5 Year
    predicted
true     1    2
   1 3970 185
   2  537 418
        auc      mmce       acc
0.7894861 0.1412916 0.8587084

 kNN - 5 Year
    predicted
true     1    2
   1 4072  83
   2  401 554
        auc       mmce        acc
0.93842793 0.09471624 0.90528376
```

```
RF - 5 Year
    predicted
true     1    2
   1 4068  87
   2  426 529
        auc      mmce       acc
0.9366480 0.1003914 0.8996086

 SVM - 5 Year
    predicted
true     1    2
   1 4084  71
   2  680 275
        auc      mmce       acc
0.8380224 0.1469667 0.8530333
```

Observation

kNN produced the best result with an AUC of 93.84%. However, Random Forest was extremely close with an AUC of 93.66%. A small difference of 0.18% is unlikely to be statistically significant. Considerations must be made between the cost of a false positive and the benefit of a true positive. In the case of Attrition, the benefit of having a true positive outweigh the cost of a false positive. The kNN curve was the best model as it has a better true positive rate than the Random Forest model.

ROC Curve – Employee turnover < 2 years




```
NB - 2 Years
    predicted
true     1   2
   1 1908 287
   2  356 349
      auc      mmce       acc
0.7866276 0.2217241 0.7782759

 RPart - 2 Years
    predicted
true     1   2
   1 2052 143
   2  363 342
      auc      mmce       acc
0.8022220 0.1744828 0.8255172

 kNN - 2 Years
    predicted
true     1   2
   1 2145  50
   2  361 344
      auc      mmce       acc
0.8689704 0.1417241 0.8582759
```

```
RF - 2 Years
    predicted
true     1   2
   1 2148  47
   2  262 443
      auc      mmce       acc
0.9115598 0.1065517 0.8934483

 SVM - 2 Years
    predicted
true     1   2
   1 2093 102
   2  444 261
      auc      mmce       acc
0.8261516 0.1882759 0.8117241
```

Observation
Random Forest produced the best result with an AUC of 91.16%.

6. Conclusion

The results show that using advanced techniques can result in highly accurate models for predicting turnover. For each model, a different level of accuracy was achieved. The smaller populations dropped in accuracy, but that is to be expected with such a limited dataset and the impact of removing more observations as the <5 year and <2 year populations were used.

The impacts of this sort of analysis, when applied to "real" datasets of employee information, can help leaders decide where to focus efforts and resources to reduce the financial and operational impacts caused by employee turnover.

For example, using this dataset and some common turnover cost estimates, the cost of turnover to the company in this dataset is over $5.4M:

Job Level	Average Annual Income	Cost factor	Avg Cost per Attrition	Attrition Count	Total Cost for Attrition
1	$ 31,178	$ 7,000	$ 7,000	143	$ 1,001,000
2	$ 69,117	35%	$ 24,191	52	$ 1,257,938
3	$ 112,661	50%	$ 56,330	32	$ 1,802,574
4	$ 157,805	60%	$ 94,683	5	$ 473,414
5	$ 233,566	75%	$ 175,174	5	$ 875,871
			Total	**237**	**$ 5,410,797**

That same analysis for employees leaving with less than 5 years of service and less than 2 years of service shows $1.95M and $1.01M in costs respectively. And remember that the cost model for turnover needs to be created and validated by each company undertaking such an analysis to ensure support and understanding of and for the analysis.

As company leaders look for ways to reduce the costs and lost productivity associated with attrition, the need to understand the true drivers of that attrition so they can form solutions to address those drivers. As reported in the feature selection section, the chi squared algorithm was used to understand what the drivers were for each of the populations examined. Looking at our population of employees who left with less than two years of service, we can say with a high level of confidence that by focusing on the following items, this company should be able to achieve a reduction in turnover for this population:

- Reduce the overtime requested employees
- Investigate further what about the job roles for these employees is driving dissatisfaction
- Consider increasing stock options

However, one must remember that these types of decision are not always easy. In this case, the populations that leave under 2 years and 5 years appear to value stock options more than the rest of the population, but there may be a reason why it matters less to the rest of the population. Decisions that change programs for all employees must be weighed for their benefit to a smaller population vs. the cost of providing that program enhancement to all employees.

This points to the purpose of undertaking such analysis and leveraging concepts like machine learning in the first place. All of this leads to better information for business leaders to make better decisions. But in the end, the business leader still must make the final decision on what actions to take. Hopefully, this type of analysis helps those decisions be better decisions.

Bibliography

Bersin, J. (2013, August 16). Employee Retention Now a Big Issue: Why the Tide has Turned. Retrieved May 30, 2019, from https://www.linkedin.com/pulse/20130816200159-131079-employee-retention-now-a-big-issue-why-the-tide-has-turned

Bureau of Labor Statistics, Employment Situation Summary. (2019, June 07). Retrieved June 08, 2019, from https://www.bls.gov/news.release/empsit.nr0.htm

Fortin, D. (2017, July 17). How to calculate employee turnover cost. Retrieved from https://www.predictiveindex.com/blog/how-to-calculate-employee-turnover-cost/

Michaels, E., Axelrod, B., & Handfield-Jones, H. (2001). The war for talent. Boston, MA: Harvard Business School Press.

Morrell, K. (2014). Understanding and measuring employee turnover. Research Handbook on Employee Turnover, 26-58. doi:10.4337/9781784711153.00007

O'Connell and Kung 2007 O'Connell, M., and M. Kung. 2007. "The Cost of Employee Turnover." Industrial Management 49 (1): 14–19