CSDA1050 Advanced Analytics Capstone Course

Presentation

Utilizing Machine Learning in Employee Retention in an Advertising Agency

Eugene Yong Geun Park

August 27th, 2019

Introduction

Advertising's main purpose is to draw human attention and interest to promote goods, services and/or messages. Marketing concepts are developed and they are delivered to public through various media means. As much as it is important to have people who can develop creative and effective marketing ideas, the business also requires people who can execute the idea in market by developing partnership and trust with client.

While there definitely is a pool of human resources with proven experience and well-known career milestones out in the market, recruitment from the pool is not always guaranteed solution for success. Availability, Suitability and Financial Feasibility are some of many factors that need to be considered for new hires and it is always very difficult to align them to current business situation.

As a result, it inevitably becomes very important to retain current Human Resources who are nurtured and trained to existing business model. Cost of losing existing resource incurs in many areas. Most immediately, the business could suffer damage in client relationship. Furthermore, there is always direct cost to replacing lost resources. It can cause substantial damage to business. No one in the industry is immune to employee retention problem. In this research, machine learning models are developed to help with this problem.

Research Question

- Can predictive machine learning models be developed to aid employee retention problem?
- Can these models help identify factors that impact employee retention?
- Can these models help make business decisions?

Dataset

A record of employees of a small advertising agency will be used. This is a dataset which consists of 130 employees from 2005 to 2019.

- S.EMP: Unique employee number assigned by the company
- Title: Official Job Title at Work
- C.LEVEL: Job level code that is assigned by the company
- Team: Team or Department employees work in(Total 5)
- Job.Level: Total 5. Assosiate(C1), Supervisor(C2, C3), Manager(C3, C4), Sr. Manager(C4, C5), Representative Officer(C6)
- BEGIN.SALARY: Beginning Salary
- Raise: Number of raises received
- H.YEAR: Hire Year
- Termination: Termination Date
- T.Year: Termination Year
- TermType: Termination by type. Quit(Voluntary Leave), Terminate(Involuntary Leave), Active
- Status: Employment status. Terminated or Active.
- DOB: Date of Birth
- Sex: Gender
- Education: Level of education. College Diploma, Bachelors Degree or Maters Degree
- Major: Field of Study
- DistToWork: Commuting distance in km

```
##     S.EMP.                   Title C.LEVEL              Team  Job.Level
## 1 60060523     Senior Art Director      C3 Creative Services Supervisor
## 2 60071662 Asso. Creative Director      C3 Creative Services Supervisor
## 3 60072318   Sr. Account Executive      C2  Account Services Supervisor
## 4 60072838   Engagement Supervisor      C3  Account Services Supervisor
## 5 60081603     Junior Art Director      C2        Production  Associate
## 6 60114364     Account Coordinator      C1  Account Services  Associate
##   BEGIN.SALARY Raise     Hire H.Year Termination T.Year TermType
## 1        70000     1 20060821   2006    20110831   2011     Quit
## 2        80000     2 20070522   2007    20110520   2011     Quit
## 3        70000     1 20070816   2007    20120323   2012     Quit
## 4        75000  -    20070910   2007    20120120   2010     Quit
## 5        35000     2 20080520   2008    20140919   2014     Quit
## 6        35000     2 20110501   2011    20140314   2014     Quit
##       Status      DOB SEX                Education
## 1 Terminated 19740928   M          Bachelors Degree
## 2 Terminated 19660123   F          Bachelors Degree
## 3 Terminated 19800925   M          Bachelors Degree
## 4 Terminated 19731212   F          Bachelors Degree
## 5 Terminated 19811224   M          Bachelors Degree
## 6 Terminated 19830126   F College/Diploma/Associate
##                          Major DistToWork
## 1 Advertising & Graphic Design       15.0
## 2           English & Religion       40.0
## 3                                    25.0
## 4                   Sociology       20.0
## 5               Graphic Design       13.3
## 6                                    24.5
```

Methodology

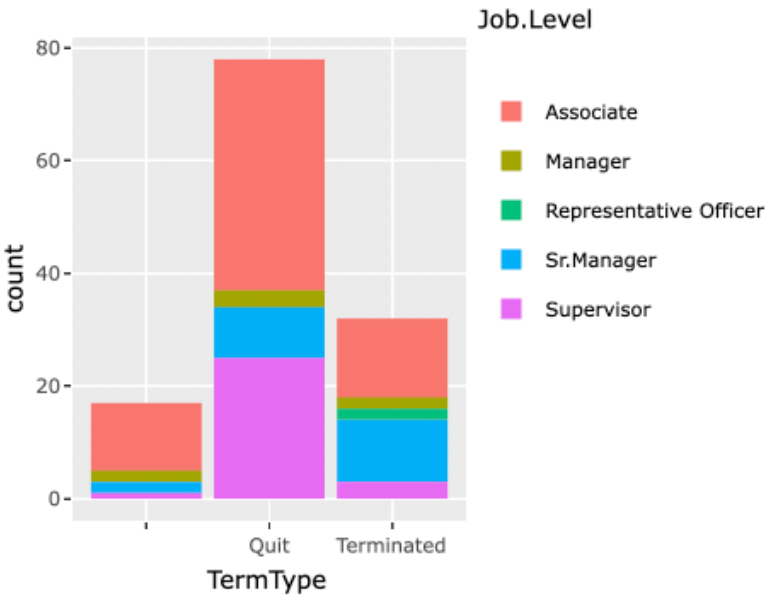
- The entire research is done in R language in R Studio.
- Data Visualization is used to gain intuitive sense around factors that impact employee retention.
- Decision Tree technique is used to build a model to classify employee's employment status (Active, Voluntary leave and Involuntary leave). Variables that are fed in the model for most optimal result are based on the learnings from Data Visualization.
- Logistic Regression model is used to build a model to predict hiring cost based on business revenue level. Correlation between recruiting cost and revenue is visually tested.

Discovery

Data Visualization

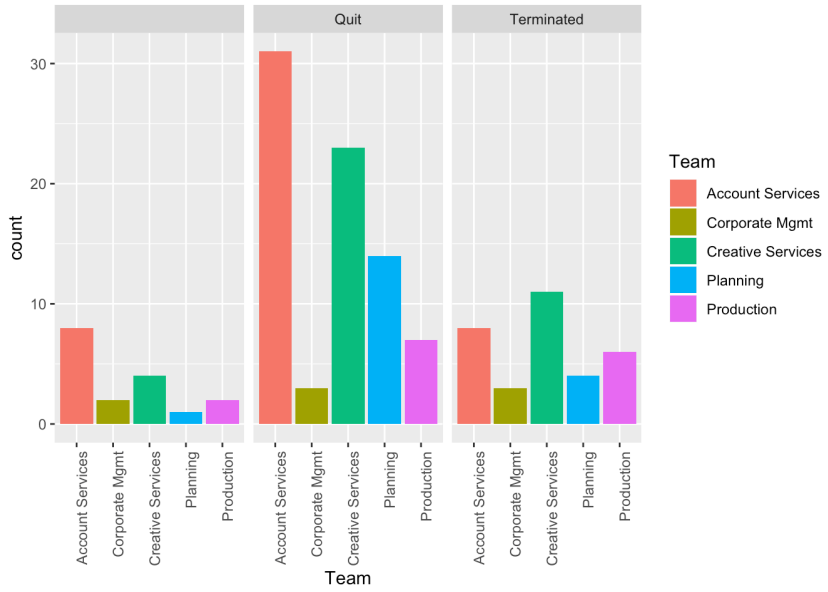
Variables from the dataset are visually examined to discover any trend in employee retention.

Employee's retention based on Job Level
Employment status is classified by job level. This suggests that associate and supervisor are the biggest groups who leaves the company voluntarily. Sr.Manager is the only group who gets terminated more frequently, instead of leaving voluntarily. This reveals some great efficiency about current operation. The management is failing to retain great resources in junior positions then chooses to terminate senior positions who are most receiving 6-figure salaries.



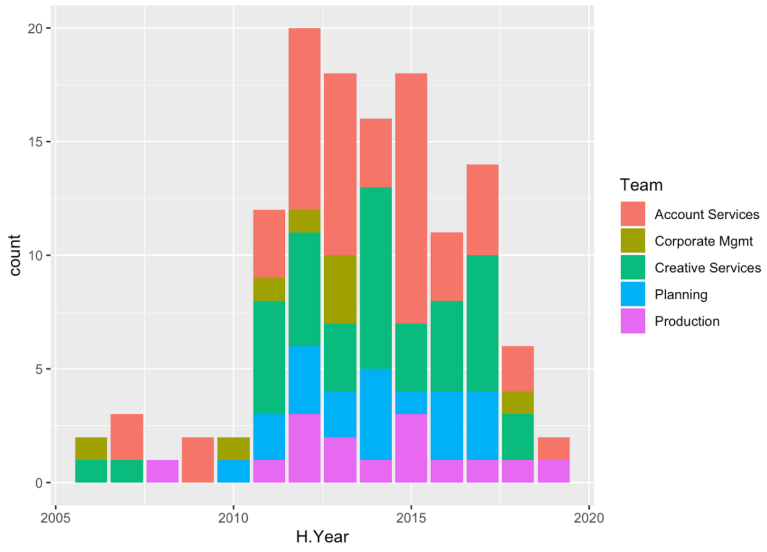
Employee's retention based on Department

Among 5 different departments in this company, this suggests that Account Services, Creative Services and Planning are the biggest groups who leave the company voluntarily. On the other hand, Corporate Management and Production departments are relatively equal in voluntary and involuntary leave. This indicates the nature of advertising industry. Account Services(Client relation) and Creative Services(Creative Concept Development) are most often recruited resources in the industry. They are constantly looking for better opportunity and being offered one from other companies.



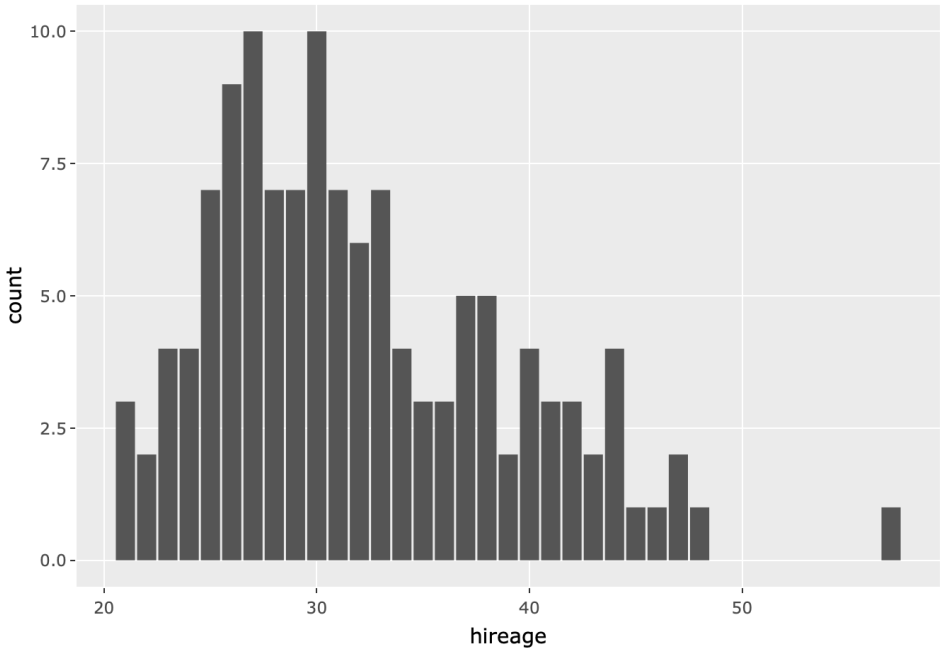
Total Number of hires by Department in each year

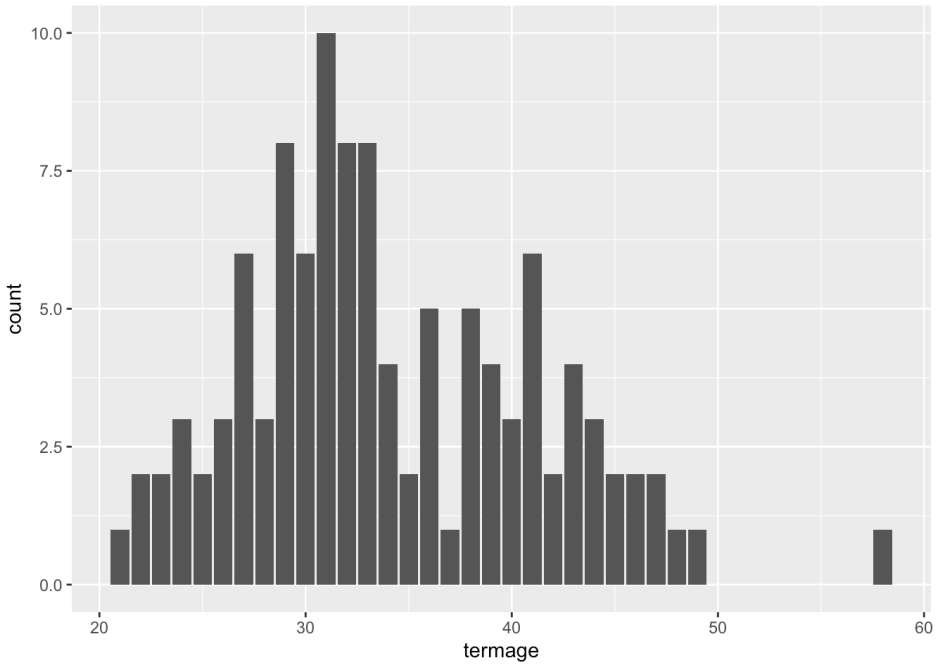
This shows that the company is in constant need of Account Services and Creative Services in most years, which is aligned to above finding.



Employee's hiring age and termination age

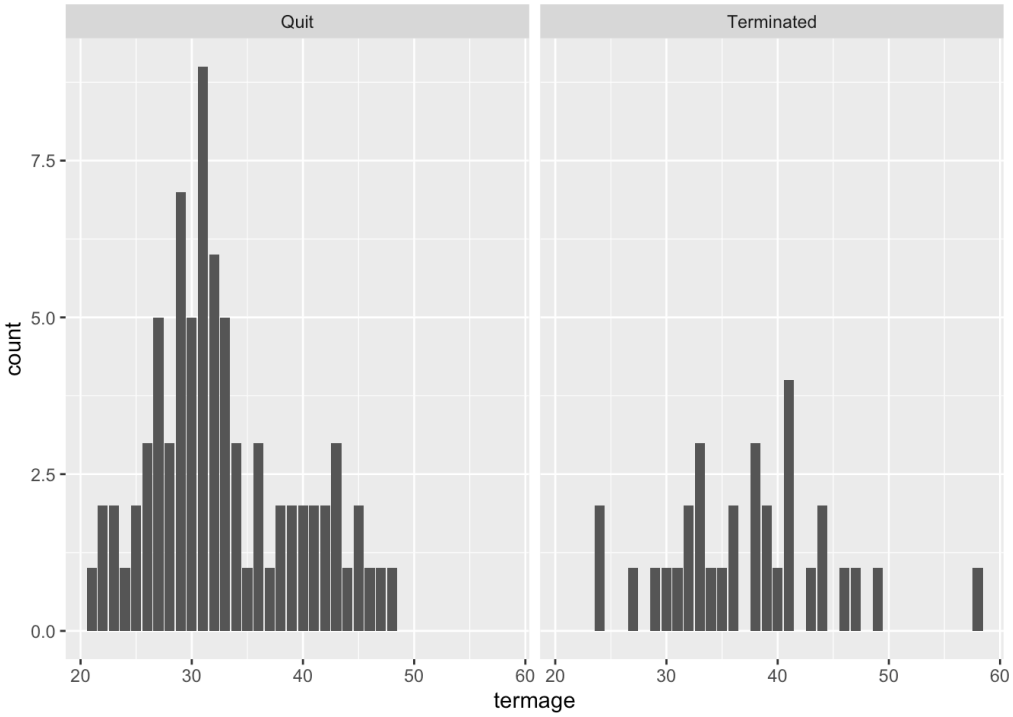
Age does not directly correlate to employee's work experience and skillset. However, there still is value in discovering age groups more frequently hired and terminated. Although below two plots look very similar, it is noticed that age group younger or equal to 30 is most actively hired, then age group in early 30s most frequently leaves the company. This may suggest that employees in younger group, under 30, levers the experience in this company to make a move to another company.





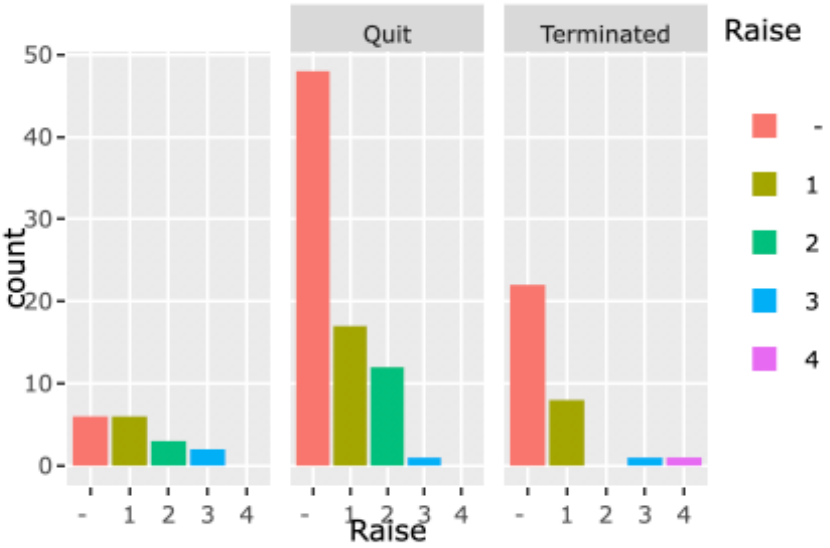
Termination age by type of termination

This again illustrates above discussed topic where in this company more junior resources chooses to quit and more senior resources are facing termination.



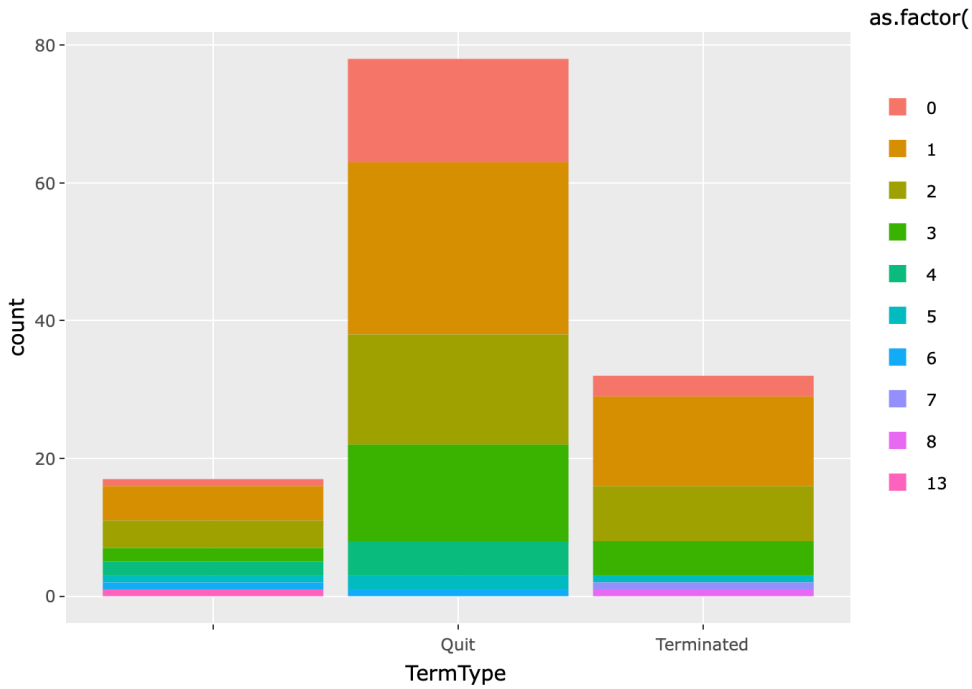
Employee's retention based on # of raised received.

This plot illustrates number of employees by number of raises received throughout their tenure. It is noticed that people with less raises choose to leave the company. Unlike other objective variables examined above, this is an objective variable and this illustrates how recognition at work place impacts employee retention.



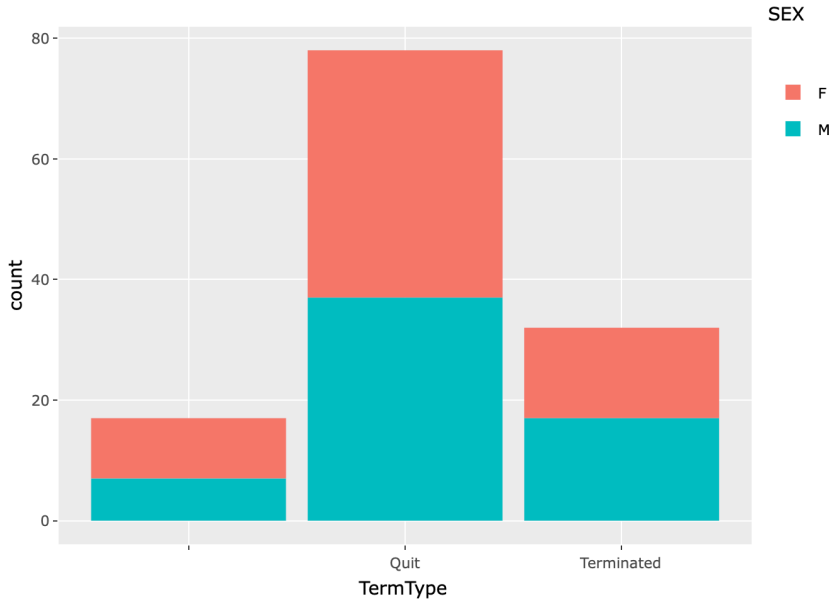
Employment Status by Tenure

Employees status is plotted by number of years they work at this company. Most notable range of tenure is between 0~4 years which employees chooses to leave the most. With 0 being exception, this coincides with the idea of people often working 1~4 years to revamp their resumes then making a jump to another organization.



Employment Status by Gender

There is not a particular pattern that influences employee retention by Gender. Although this does not help the research, this illustrates that this workplace is not gender discriminatory.



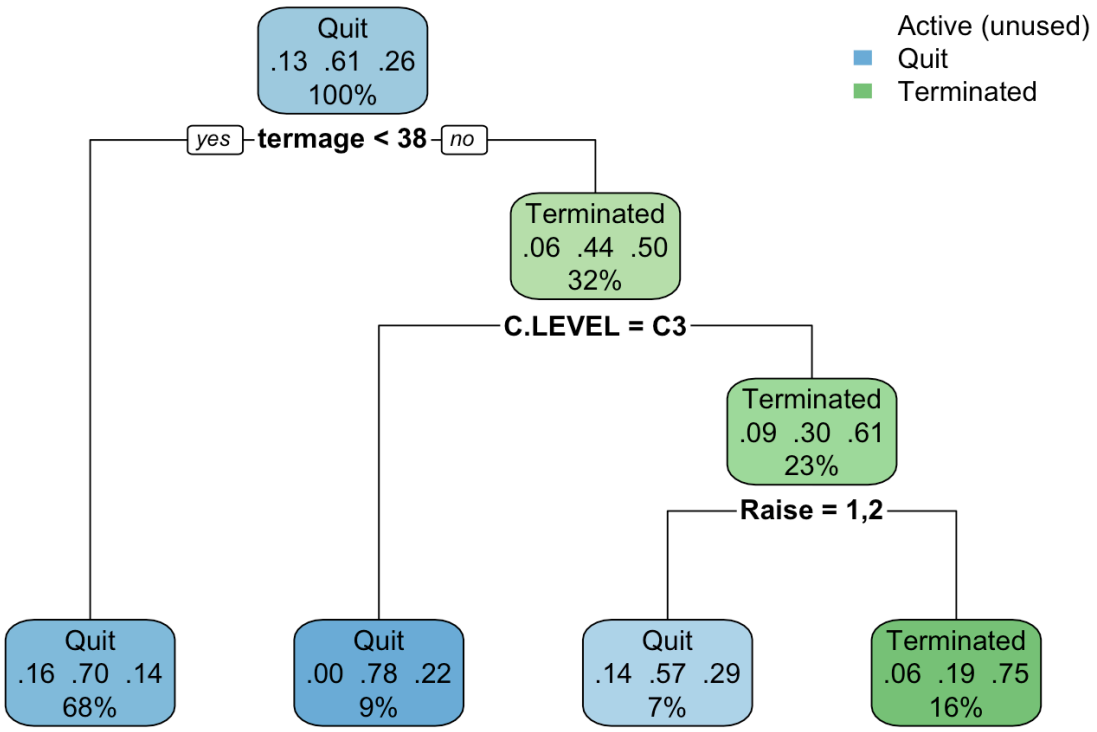
Predictive Modelling

After exploring the dataset through data visualization, machine learning techniques are used to build predictive model. Decision Tree technique is used to develop a model that can predict employees at flight risk and identify crucial variables that influence the decision. Then Linear Regression is used to predict hiring cost based on revenue. The dataset is subset by 80% train and 20% test.

Decision Tree using all variables examined in Data Visualization phase

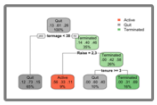
This Decision Tree model causes two problems. First of all, none of the nodes can have Active employee are majority, therefore it fails to develop a criterion that can predict Active status. Secondly, the overall accuracy is only 53.85%.

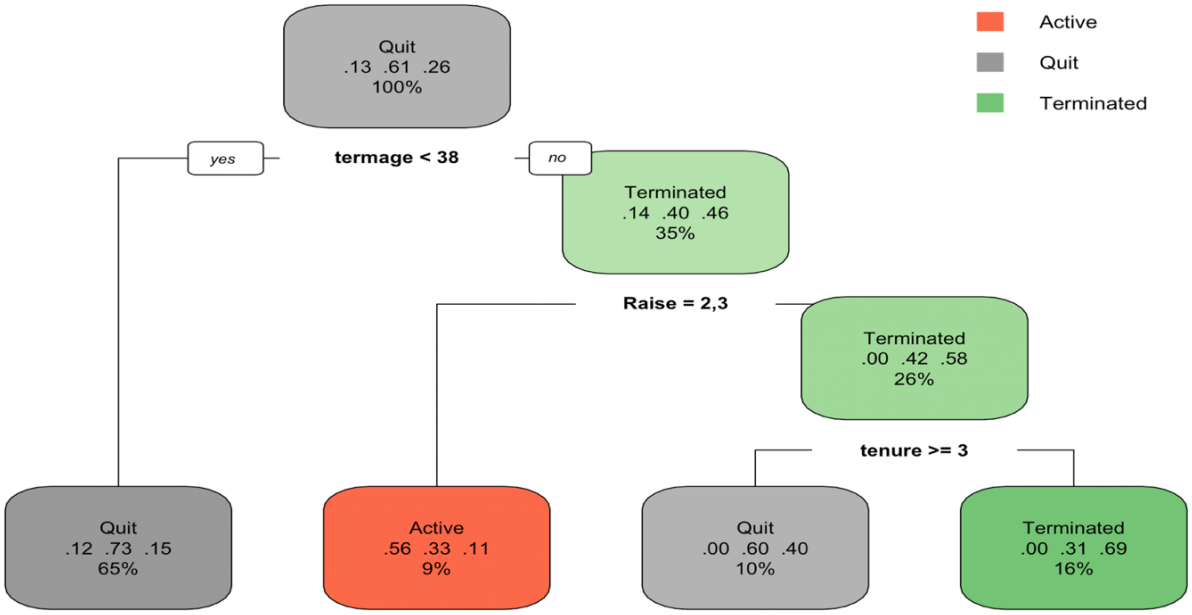
```
fit <- rpart(TermType~., data=train)
rpart.plot(fit)
```



Decision Tree using Termination Age, Number of Raises, and Tenure data

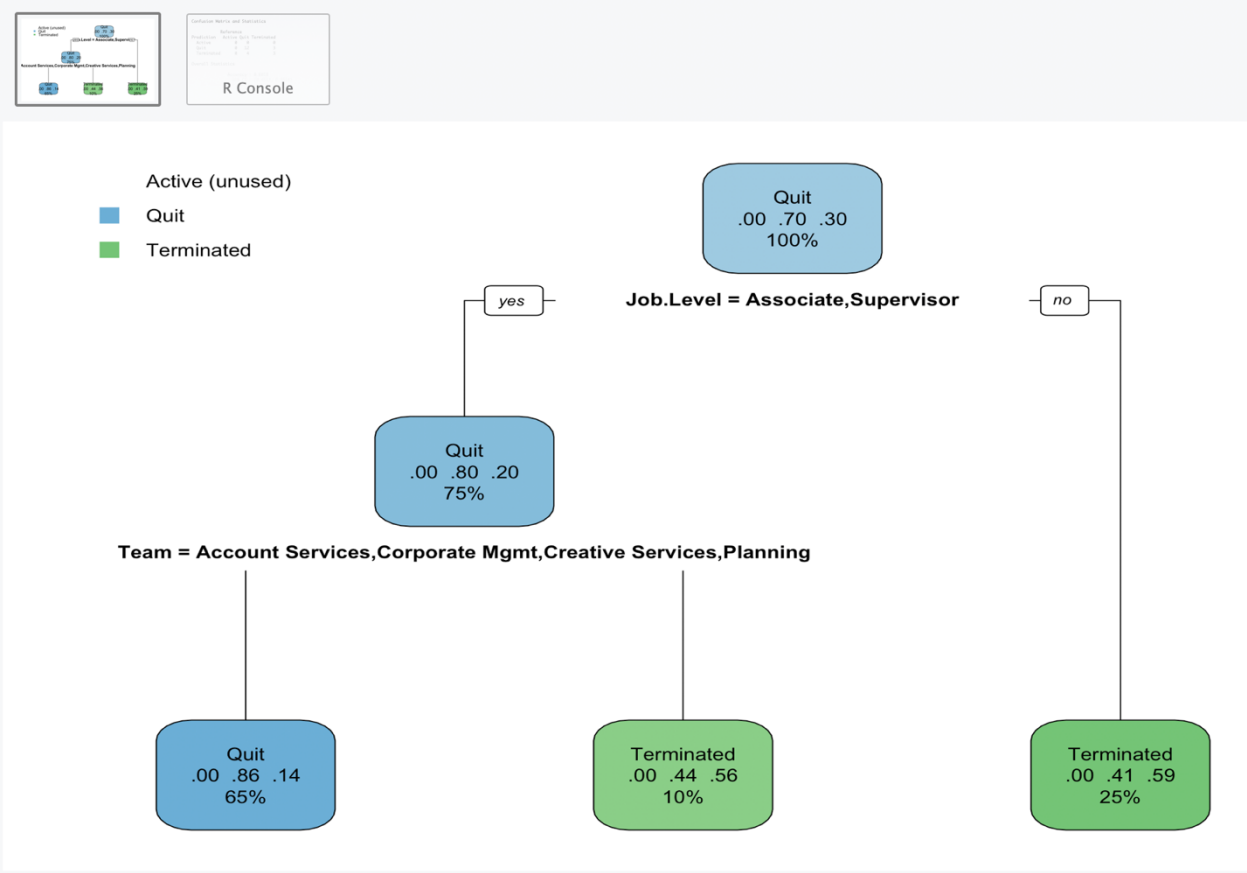
Below model is built with Termination Age, Number of Raises received and Tenure. This model has success in identifying all three different employee status (Active, Voluntary Leave and Involuntary Leave). The criterion developed for Active status seems reasonable, which while being above age 38, recognition (Raises 2 or 3 times) is received. However, the overall accuracy of the model is gotten even lower than first developed model, 46.15%.





Decision Tree using Job Level and Department data

Among many tried, this Decision Tree model below has the highest accuracy, 68.18%. Variables chosen to build this tree coincide with what are mostly visually and intuitively noticed in Data Visualization (Job Level and Department).



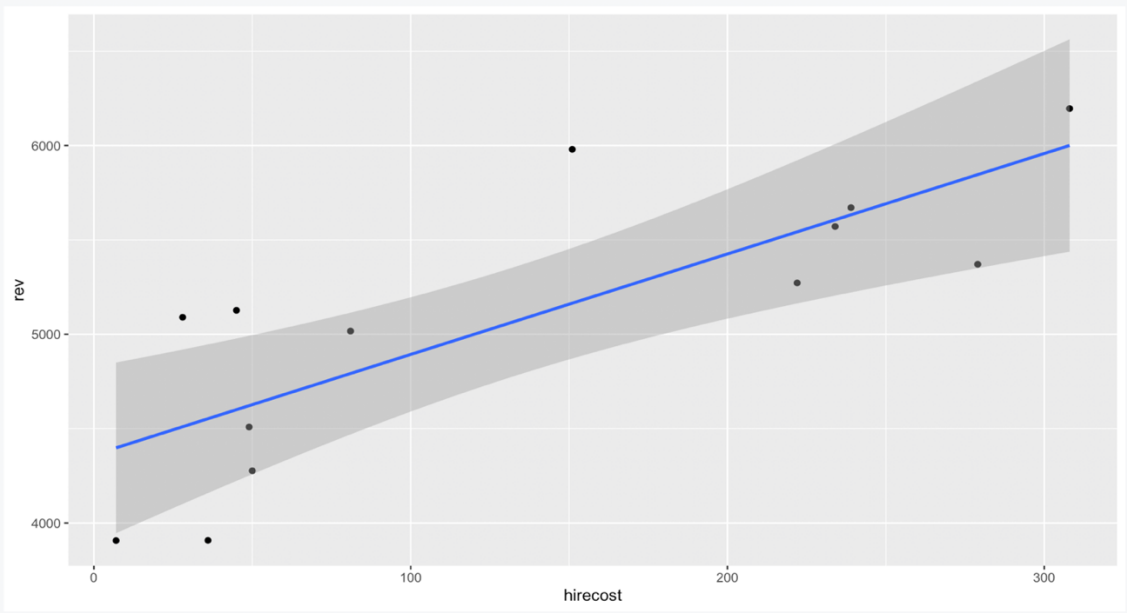
Again, none of the decision tree models produce great accuracy, all under 70%. However, the range of variables tested to improve their performance match what are intuitively and visually considered important in data exploration stage.

Linear Regression

As decision tree does not produce reliable predictive model with great accuracy, linear regression is considered to add business value in different perspective. If an employee is hired through a recruiting firm, commission paid for the recruitment is roughly 22% of beginning salary. As all employees beginning salary is available, their estimated hiring(recruiting) cost can be calculated. Then the hiring cost is aggregated by year for annual total. When the data is combined with annual revenue, then a dataset is created to fit into a linear regression model to examine correlation.

```
##    year hirecost hirecount  rev avgcost
## 1  2006       36          2 3908      18
## 2  2007       49          3 4509      16
## 3  2008        7          1 3907       7
## 4  2009       50          2 4277      25
## 5  2010       45          2 5127      22
## 6  2011      234         12 5571      19
## 7  2012      279         20 5371      13
## 8  2013      239         18 5671      13
## 9  2014      308         16 6196      19
## 11 2016      151         11 5980      13
## 12 2017      222         14 5272      15
## 13 2018       81          6 5017      13
## 14 2019       28          2 5090      14
```

- Year: The year for each record
- hirecost: Hiring Cost or Recruiting Cost, in thousands
- hirecount: Number of hires made in the year
- rev: Revenue, in thousands
- avgcost: Average cost per hire



When plotted, although not perfect, it visually demonstrates a correlation between hiring cost and revenue. Based on this, a linear regression model is developed.

```
Call:
lm(formula = hirecost ~ rev, data = hire)

Residuals:
     Min       1Q   Median       3Q      Max
-107.459  -45.942    9.566   42.421  110.755

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -458.41206  140.66161  -3.259  0.00761 **
rev            0.11667    0.02748   4.245  0.00138 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 70.03 on 11 degrees of freedom
Multiple R-squared:  0.621,     Adjusted R-squared:  0.5865
F-statistic: 18.02 on 1 and 11 DF,  p-value: 0.001377
```

The linear regression model establishes correlation between hiring cost and business revenue numerically, which is expressed in formula, Estimated Hiring Cost= Revenue * 0.11667 - 458.41206.

Conclusion

When reflecting back to stated three research questions,

- Can predictive machine learning models be developed to aid employee retention problem?
- Can these models help identify factors that impact employee retention?
- Can these models help make business decisions?

It is most certain that machine learning can add value to current business. Decision Tree models have suggested the area that has most significance in influencing employee retention in this company. There clearly is a trend that employees in junior role and particular department have higher probability of leaving the company voluntarily, meanwhile senior positions are more often faced involuntary terminations. Moreover, linear regression model identifies a correlation between hiring cost and business revenue. This correlation has produced a formula that can estimate hiring cost based on expected business revenue.

The information gained from examined models can be used in actual business operation. For instance, when annual financial planning takes a place, the year’s hiring cost can be estimated based the target revenue. Then current employees’ status can be reviewed to identify what tested Decision Tree model has suggested as at-risk prospects. Given the estimated hiring cost, it would make sense to use the budget to pro-actively protect existing resources, instead of re-actively using the budget to make new hires.

This does not suggest that developed models are highly reliable solutions to the matter. Developed Decision Tree models do not perform in great accuracy. There is a good change that estimated cost from this Linear Regression will significantly fluctuate in real business situation. Most of all, the size of dataset used is too small to test and train a proper model. Despite incompleteness of developed model, utilizing machine learning is trusted to be the right path to take in business management perspective. For many years, managing Human Resources in this company has largely relied on intuition and maybe that is what limited the organization to grow faster. As it is understood that this approach can add value, there should be increased effort in maintaining quality database in all business areas and continued investment in data analytic tools and its users.