CSDA1050 Advanced Analytics Capstone Course

Final Report

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. Age, gender, address, hire date, termination date, termination reason, starting salary, department, job title, job level, commuting distance and education are currently available. Using available data, hiring age, termination age and tenure can also be created. The biggest challenge with the dataset will be its size. Its small size may not result in a reliable model.

Methodology

- The entire research is done using 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.

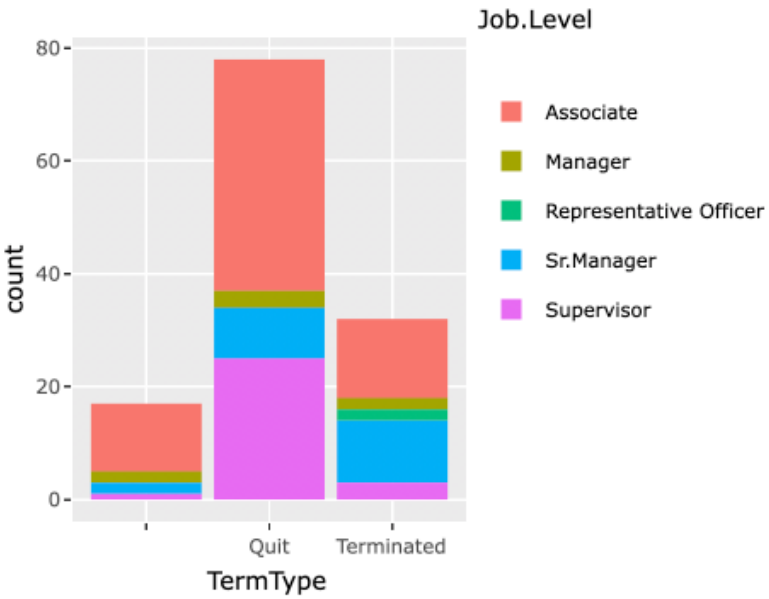
Discovery

Data Visualization

Variables from the dataset are visually examined to discover any trend in employee retention. Among many below three seem to show stronger trend than others.

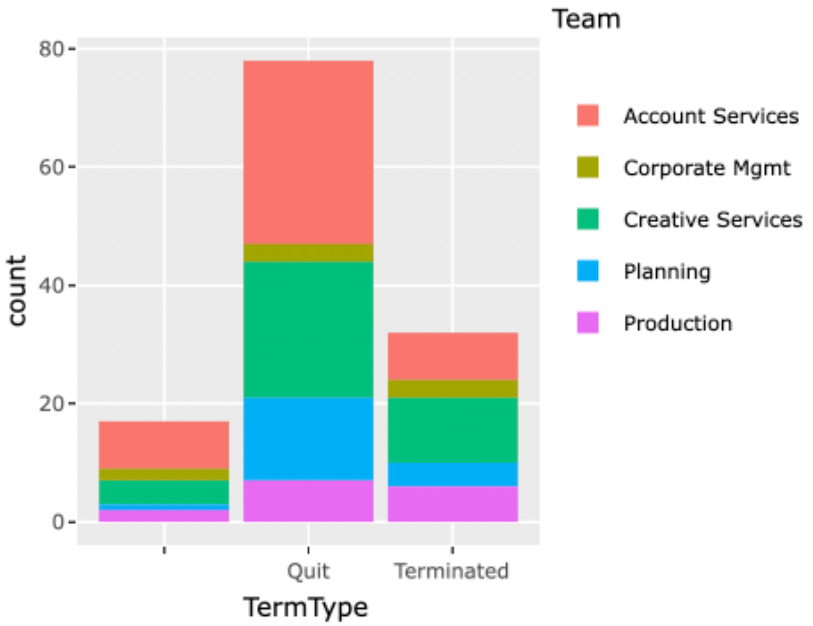
Employee's retention based on Job Level

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.



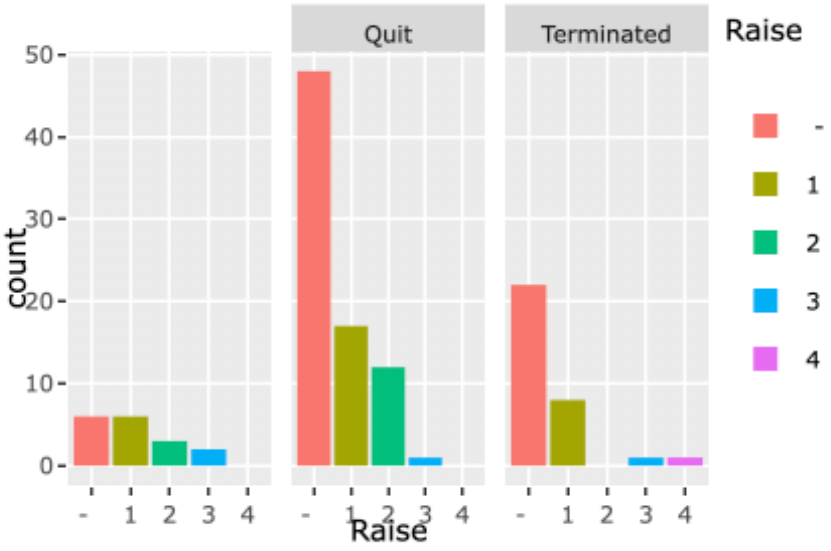
Employee's retention based on Department

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.



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. It suggests how recognition at work place impacts employee retention.

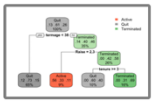


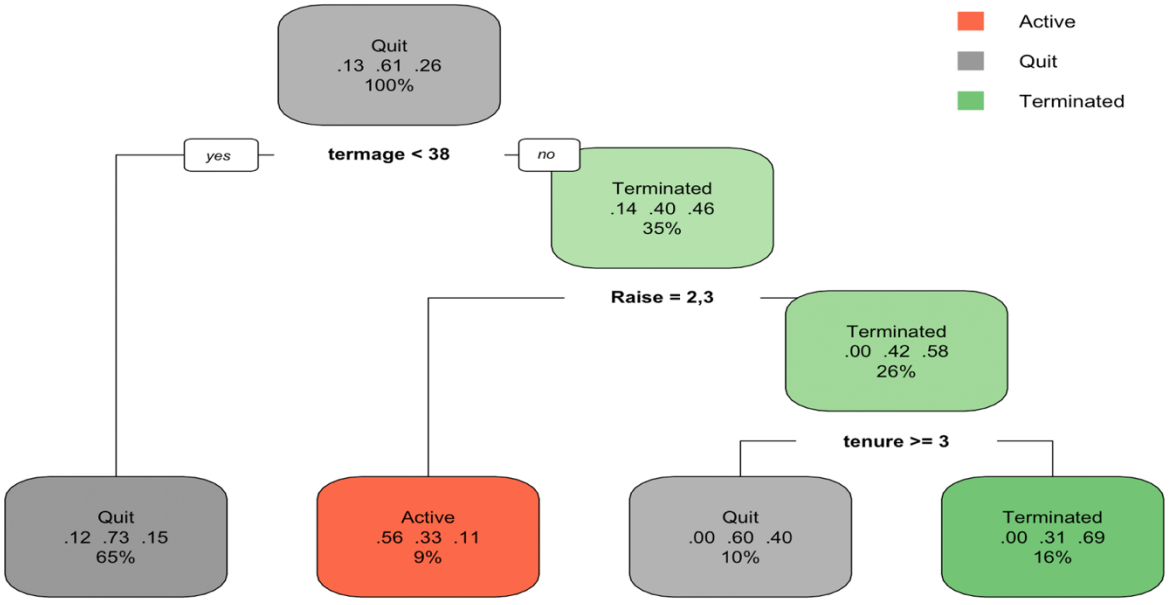
Predictive Modelling

After exploring the dataset through data visualization, machine learning techniques are used to build predictive model.

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

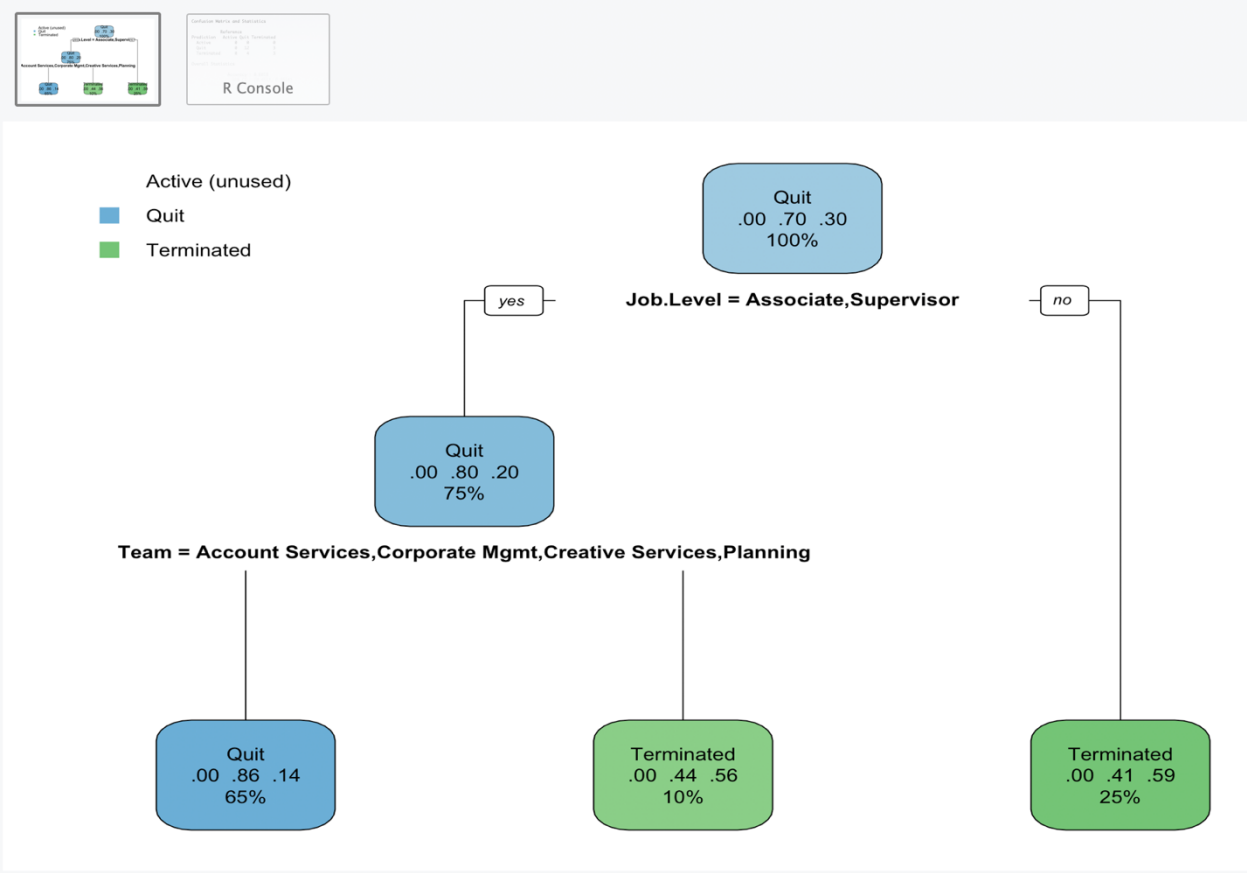
Decision Tree technique is used to build a predictive model that can identify employee's retention status based on variables. Below model is built with Termination Age, Number of Raises received and Tenure. This model has success in identifying all three different employee status (Active, Voluntary Leave and Involuntary Leave). However, the overall accuracy of the model is not very high, 46.15%. This is a demonstration of importance in variable selection when building a decision model.





Decision Tree using Job Level and Department data

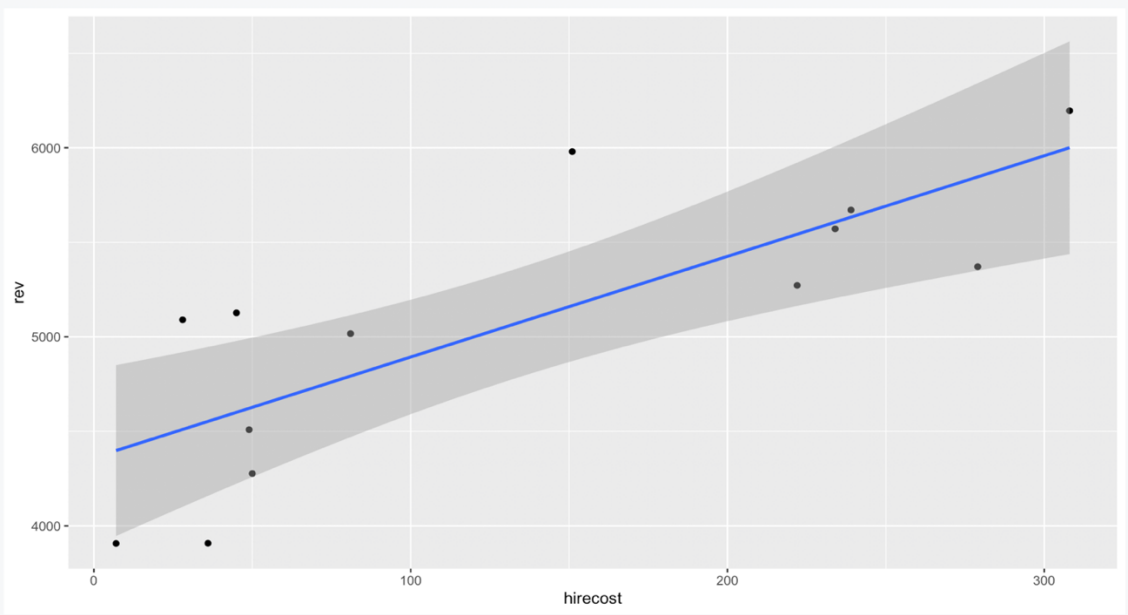
Among many tried, this Decision Tree model below has the highest accuracy, 68.18%. Variables chosen to build this tree coincide with what are mostly visually and intuitively noticed in Data Visualization (Job Level and Department).



Neither of the accuracy from above two models are reliable (46.15% and 68.18%). While this probably is due to small size of dataset, it is still insightful to see where, when and based on what the nodes branch out.

Linear Regression

In addition to Decision Tree models that can predict employment status, a linear regression model is built to predict hiring cost based on level of business revenue. A Correlation between these features are visibly noticed.



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.

```
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
```

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.