ST308 Coursework

HR Analytics: A Bayesian analysis of attrition tendencies of young-to-mid career new employee hires

Table of Contents

1. Introduction

Attrition is a foremost concern for all organizations across sectors. A corollary of the dissipation of knowledge and intellect, a disruption to project pipelines, and the need for channelling resources into talent recruitment, attrition has a major impact on a company's operational efficiency, costs and long-term health (O'Connor, 2020). Human resources literature between 1955 and 2014 revealed a variety of factors affecting an employee's decision on attrition, including dissatisfaction with compensation, inadequacy in the internal and external equity, toxic work environment, incompatible work culture, employee attitude, insufficient support, unsatisfactory work relationships, inadequate opportunities for growth, hiring practices and managerial style (Swathi, 2019).

In 2018, the technology sector has the highest average turnover rate of 13.2%, followed by 13.0% for the retail and consumer products sector, and 11.4% for the media and entertainment sector (Booz, 2018). As labour resource becomes increasingly competitive, human resource functions of organizations have been closely involved with the capture, measurement, and organization of workforce data to improve talent attraction and retention using data analytics. For example, IBM Watson, a healthcare industry leader which faces an estimated attrition rate between 14.3% to 18% (Singh, 2020), has developed a "predictive attrition programme" using Artificial Intelligence to predict whether an employee is likely to leave with a 95% accuracy rate (McLaren, 2019).

With the salience of HR analytics, this project aims to streamline and improve talent management operations for an anonymised company in the healthcare sector by predicting attrition tendencies and assessing the linkages between company policies and attrition tendencies, with a focus on Bayesian methods. The company, which has a reported attrition rate of 15% in 2018, has underperformed in its talent retention policies relative to the industry average of 9.4% attrition rate in the pharmaceutical and healthcare sector (Booz, 2018). This project is divided into two segments: (i) Selection of Bayesian logistic regression models to improve the prediction of attrition tendencies of new young-to-mid-career company hires, and (ii) Assessment of company-related policies affecting attrition of new young-to-mid-career company hires.

Given the multitude of potential factors affecting attrition, this analysis is highly relevant for the company as it helps the company to forecast and pre-empt future attrition, and suggests tailored policy solutions to improve talent retention outcomes. A Bayesian approach is critical for both parts of the project, where the specification of priors allow for the construction of regularised models for attrition prediction in part I, while the flexibility of Bayesian models allows for a more nuanced understanding of the linkages between company policies and attrition decisions.

2. Data description and preparation

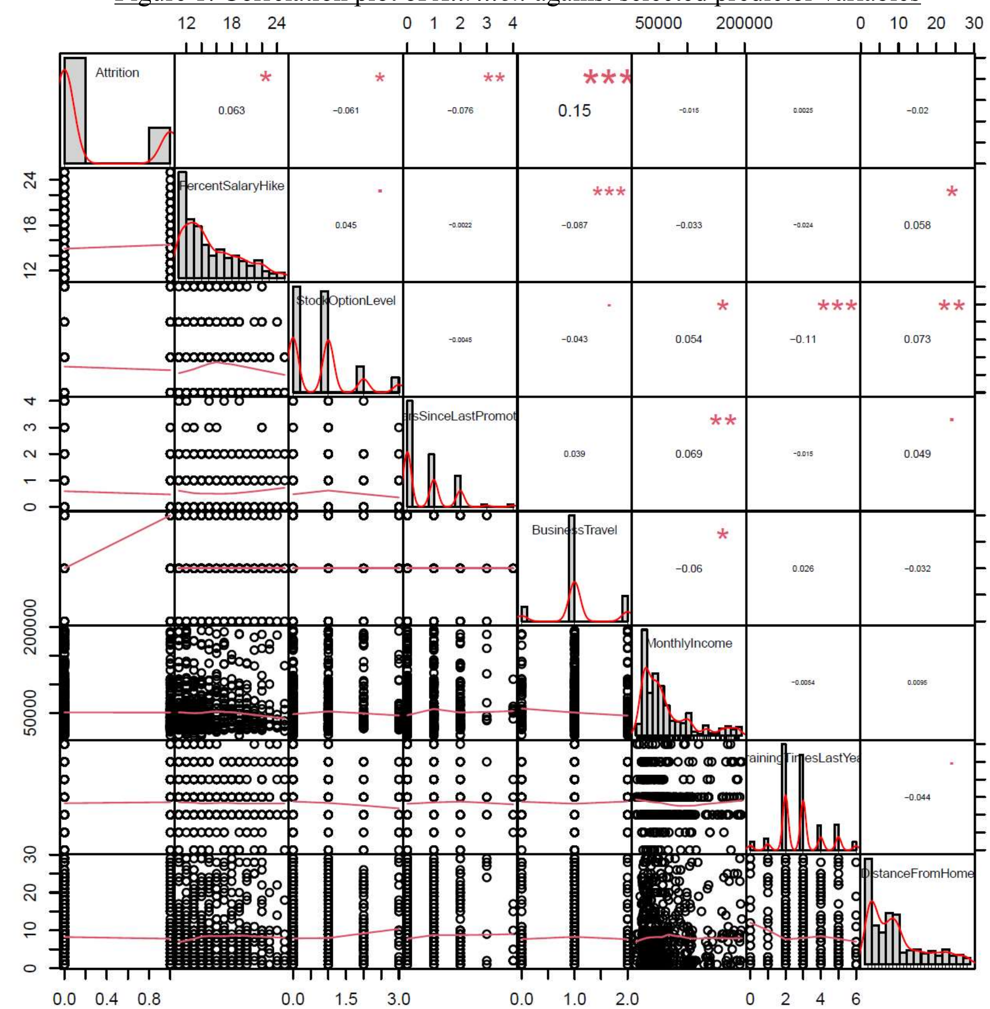
I conducted my data analysis employee data from the aforementioned anonymised company from the healthcare sector, obtained from Kaggle as of 7th April 2021. The original dataset consists of 4,411 observations of unique employees, with 24 columns and no missing observations.

Both parts of my analysis pertain to young-to-mid-career new company hires. To select young-to-mid-career professionals, I restricted the data to individuals with less than 40 years of age for young professionals. For the remaining individuals with age ranging from 18 to 39 years old, I created 11 *AgeGroup(s)* for every 2 consecutive ages. I further restricted the data to individuals with less than or equals to 5 years with the company to restrict our attention on new hires, leaving us with 1,581 observations. I identified the existence of outliers by plotting the histograms of continuous variables and summarising the various percentiles of the variable. After identifying *MonthlyIncome* as the only variable containing outliers, I removed all observations with values above the 99th percentile of the variable to prevent distortion of results, following the capping method for outlier removal (Tiwari et al, 2007). *Attrition*, which is a binary variable, was converted into numeric values of 0 and 1's. *BusinessTravel*, which includes *Non-Travel¸ Travel-Rarely*, and *Travel-Frequently*, is converted into a numeric scale with the listed observations taking values of 0, 1 and 2 respectively. *JobRole*, which includes 9 unique observation types – namely *Human Resources, Laboratory Technician, Sales Representative, Sales Executive, Healthcare Representative, Research Scientist, Research Director, Manufacturing Director,* and *Manager* – are mapped into categorical numbers ranging from 1 to 9.

Out of the 18 variables that were included in our predictive model, I shortlisted seven variables that are deemed to be pertinent to company policies. These are: *PercentSalaryHike, StockOptionLevel, YearsSinceLastPromotion, BusinessTravel, MonthlyIncome, TrainingTimesLastYear,* and *DistanceFromHome*. The first six variables are a direct consequence of the company's policies. *DistanceFromHome*, on the other hand, has also been included as HR policies can possibly mitigate this factor should it be deemed to be important. For example, the provision or expansion of the company's shuttle bus services or the setting up of new offices can reduce the impact of long commute distance on an employee's decision to attrit.

Figure 1 consists of the distributions and correlation plots for each variable. An initial examination of the distributions indicate that outliers have been successfully removed. The correlation plots, on the other hand, seem to reveal differences in the correlation strengths of different variables, with a weak correlation observed between *Attrition* and *MonthlyIncome, TrainingTimesLastYear,* and *DistanceFromHome*.

Figure 1: Correlation plot of *Attrition* against selected predictor variables



To improve the running time of my subsequent analysis, I used feature selection techniques to further select relevant variables out of the seven selected variables. I conducted the standard logistic regression, as well as the ridge and lasso logistic regressions with the "glmnet" package in R. I supplemented my understanding on the best set of company-related features affecting *Attrition* by performing a backward stepAIC logistic regression. The stepAIC "backward" approach is a stepwise regression where variables are iteratively eliminated from the regression model based on the Ákaike Information Criterion (AIC), which is a metric that favours a model with high log-likelihood and low model complexity (Brownlee, 2020). To estimate the predictive accuracy of the above models, I split the dataset into training and testing datasets with a 7:3 proportion, randomly sampled, and trained the models using the training dataset. I then ran the model on the testing dataset to predict outcomes, which in turn allows me to calculate the mean error rate of the models to facilitate model comparison.

Table 1: Feature selection of variables pertinent to company policies

	Standard	stepAIC	Ridge	Lasso
Constant	-2.32***	-2.37***	4.60x10-2	7.05x10-2
	(0.352)	(0.302)		
PercentSalaryHike	0.0502**	0.0500**	8.69x10-3	6.33x10-3
	(0.0160)	(0.0160)		
StockOptionLevel	-0.163*	-0.162*	-2.76x10-2	-1.79x10-2
	(0.0712)	(0.0702)		
YearsSinceLastPromotion	-0.245***	-2.44***	-3.72x10-2	-2.87x10-2
	(0.0727)	(0.0724)		
BusinessTravel	0.725***	0.723***	0.123	0.112
	(0.116)	(0.116)		
MonthlyIncome	4.19x10-7	-	4.06x10-8	-
	(1.32x10-6)			
TrainingTimesLastYear	-0.0164	-	-2.32x10-3	-
	(0.0486)			
DistanceFromHome	-3.64x10-3	-	-6.22x10-4	-
	(7.75x10-3)			
Mean error rate	0.2316	0.2316	0.2379	0.2379

Note: *p<0.1; **p<0.05; ***p<0.01

Table 1 reports the coefficient estimates of the models, with the standard errors included where it is relevant. Importantly, both ridge and lasso logistic regression have a mean error rate of 0.238, while the standard logistic regression and stepAIC logistic regression have mean error rates of 0.232. As a result, I chose the four variables that have been selected from the stepAIC model to be included for my subsequent analysis. These are: *PercentSalaryHike, StockOptionLevel, YearsSinceLastPromotion,* and *BusinessTravel*. These variables correspond to those that were shortlisted from the lasso logistic regression. To account for heterogeneous treatment effects across subpopulations, I retained important categorical variables, namely the factors *AgeGroup* and *JobRole*.

3. Methodology and results

For part I of the project, I compared the performances of various Bayesian logistic regression models using the four shortlisted variables as predictors of *Attrition – PercentSalaryHike, StockOptionLevel, YearsSinceLastPromotion,* and *BusinessTravel*. For part II of the project, I assessed the heterogeneity in the associations between shortlisted factors and *Attrition* across *AgeGroup* and *JobRole* to inform nuanced policy decisions. This section aims to provide an overview of the methods and findings of this project.

Part I: Selection of Bayesian logistic regression models to improve prediction of attrition tendencies

Using the Markov Chain Monte Carlo (MCMC) method, I performed Bayesian logistic regressions with different priors and compared the models' performances. Logistic regression is a discriminative classification model that is used to perform statistical inference and predict the occurrence of a particular event, which in this case is employee attrition. I first conducted a standard logistic regression by setting vague Cauchy priors for the unknown parameters, following Gelman et al (2008). I then conducted Ridge and Lasso logistic regressions by setting Normal and Laplace prior distributions for the coefficients respectively, with mean set to zero for shrinkage, following van Erp et al. (2019). Both models can also be interpreted as adding a penalty term to the magnitude and number of coefficients in the regression model.

To allow for heterogenous effects across subpopulations, I performed hierarchical modelling using different job roles and age groups as subpopulations. Hierarchical models allow for coefficients for the same predictor variable to vary across subpopulations, but at the same time restrict the variations in coefficients across subpopulations by imposing a common distribution across the coefficients. As opposed to conducting separate regressions on various subpopulations to determine their respective coefficients, hierarchical models are more realistic for modelling heterogenous effects across subpopulations as it is expected that the same causal channel manifests across all subpopulations but only varies in intensity, hence coefficients across subpopulations should not differ substantially from each other.

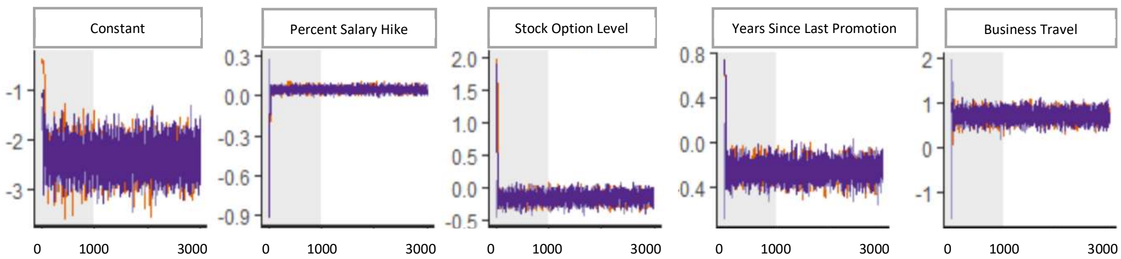
Table 2 contains the parameter estimates of the standard logistic regression, ridge regression and lasso regression models, and the population parameters of the two types of hierarchical models. The magnitude of the constant term and the coefficient estimates for *StockOptionLevel*, *YearsSinceLastPromotion*, and *BusinessTravel* are smaller in the ridge and lasso regressions compared to the standard logistic regression due to the effect of the penalty term, lambda. For both models, lambda is a tuning parameter, where a larger value for lambda is indicative of a smaller variance on the prior of the parameter estimates and hence a stronger shrinkage effect (van Erp et al., 2019).

Table 2: Parameter estimates of Bayesian logistic regression models

	Standard	Ridge	Lasso	Hierarchical (Age Group)	Hierarchical (Job Role)
Constant	-2.37	-2.27	-2.31	-2.38	-2.50
PercentSalaryHike	0.05	0.05	0.05	0.04	0.05
StockOptionLevel	-0.16	-0.16	-0.15	-0.23	-0.17
YearsSinceLastPromotion	-0.25	-0.23	-0.23	-0.38	-0.27
BusinessTravel	0.72	0.65	0.68	1.11	0.78
Lambda	-	1.44	3.51	-	-
AIC	4541.10	4515.67	4524.04	5173.41	4927.89
BIC	4567.92	4547.86	4556.24	5522.19	5223.01

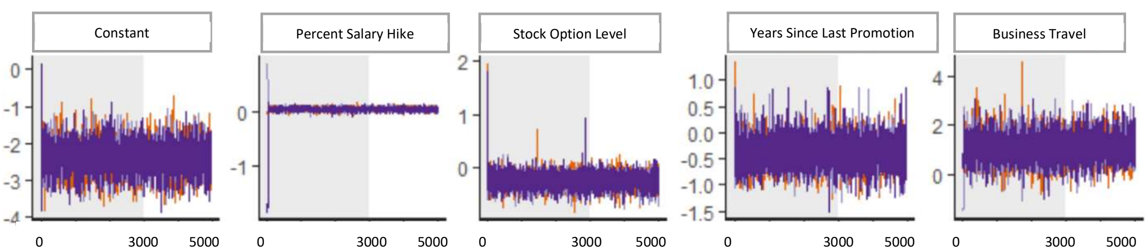
To compare the models based on fit and complexity, I calculated the AIC and the Bayesian Information Criteria (BIC) of all regression models, with smaller AIC and BIC values indicating better model performances. Both metrics penalise a model for having more free parameters and a lower log-likelihood, where the former is an indicator of higher model complexity and the latter an indicator of poorer model fit. Compared to AIC, BIC imposes a harsher penalty on model complexity (Brownlee, 2020). As reported in Table 2, the ridge regression model has the lowest AIC and BIC values. Both ridge and lasso regression models also have better model performances compared to the standard logistic regression, despite the inclusion of an additional free parameter, the penalty term. This suggests that overfitting may be a substantial issue in the standard logistic regression, and regularization in this case helped to increase the log-likelihood of the model. On the other hand, both hierarchical models have larger AIC and BIC values due to the significantly higher model complexity, where a large number of free parameters is used to measure the different coefficient estimates for each subpopulation.

Figure 2: Trace plots of parameter estimates for standard logistic regression model



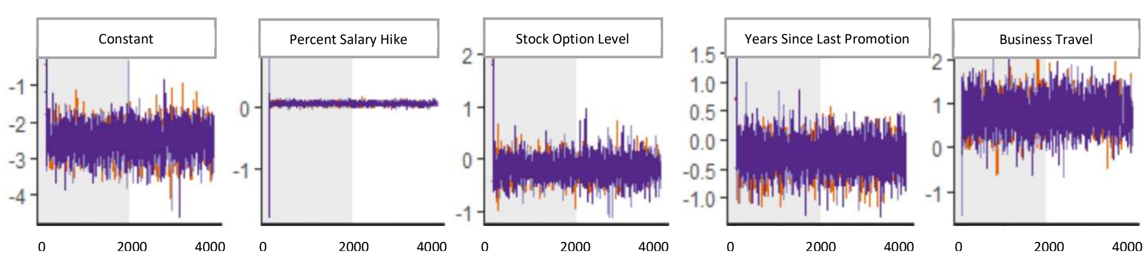
To assess the reliability of the MCMC estimates, I examined the trace plots of the Markov chains to ensure that they have converged. Figure 2 shows the trace plots of the three Markov chains of the parameter estimates of the standard logistic regression, denoted by the different colours. Markov chains from randomly initialisation points appear to have converged after the burn-in period that is shaded in grey. This is verified by the fact that Rhat value of coefficient estimates are all 1.00, showing that chains have mixed. Effective sample sizes are all above 3,000, which suggests that parameter estimates are reliable according to the Bayesian central limit theorem. Similar trace plots of the Markov chains, effective sample sizes and Rhats apply to the Ridge and Lasso logistic regressions, suggesting that the Ridge and Lasso regression estimates are also reliable.

Figure 3: Trace plots of coefficient estimates of hierarchical model (age group)



For the hierarchical model with age groups, the Markov Chain Monte Carlo estimation reported that there were 316 divergent transitions out of 6000 post-warmup samples, which is approximately 5% of the overall number of draws. Figure 3 shows the trace plots of the population coefficient estimates in the model, where it is evident that the Markov chains have mixed. This is supported by the fact that Rhat values for all coefficient estimates are less than or equal to 1.01. The effective sample sizes of all parameter estimates are mostly above 1,000, with the minimum being 478 for the between standard error of *StockOptionLevel*. The between standard error of coefficient estimates across age groups are 0.05, 0.23, 0.62, and 1.32 for *PercentSalaryHike, StockOptionLevel, YearsSinceLastPromotion,* and *BusinessTravel* respectively.

Figure 4: Trace plots of coefficient estimates of hierarchical model (job role)



For the hierarchical model for job roles, the Markov Chain Monte Carlo estimation reported that the effective sample sizes for all coefficient estimates are above 600 and Rhat = 1.00 for all parameter estimates. Figure 4 shows the trace plots of the population coefficient estimates in the model, where it is evident that the Markov chains have mixed. There are 479 divergent transitions out of 6,000 post-warmup draws, which is approximately 8% of the total draws. The between standard errors are 0.03, 0.32, 0.54 and 0.69 for coefficient estimates of *PercentSalaryHike, StockOptionLevel, YearsSinceLastPromotion,* and *BusinessTravel* respectively.

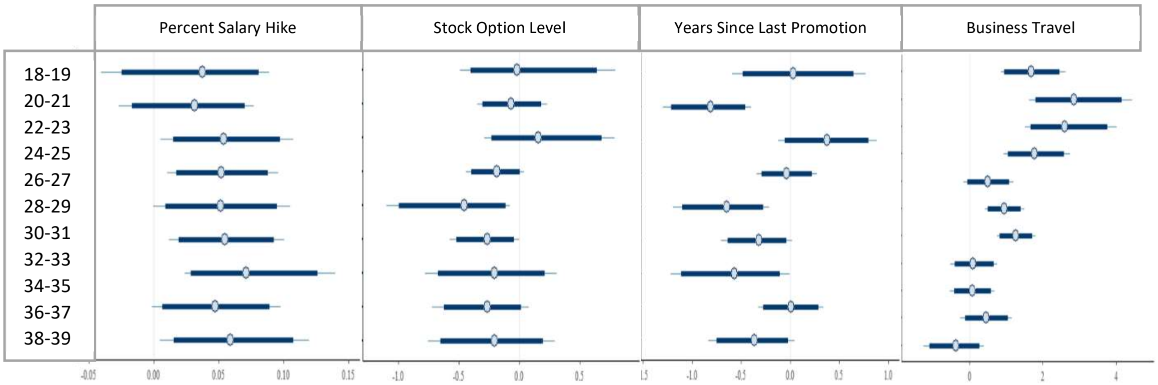
The comparison of AIC and BIC values of models indicate that ridge regression model is the best performing model for predicting attrition tendencies. Referring to the coefficient estimates of the ridge regression in Table 2, *PercentageSalaryHike*, which measures the percentage salary hike in the previous year, appears to be positively associated with the attrition of employees. Specifically, 1 percentage salary hike is associated with 0.050 units increase in log-odds, all else equals. Though this may seem counterintuitive, the positive association could be explained by the fact that employees may time their resignation after a salary hike in order to hold higher bargaining power at the new company during salary negotiation. One unit increase in *StockOptionLevel*, on the other hand, is associated with a 0.16 decrease in the log-odds of attrition, all else equals. This is reasonable as an employee with a greater share of stocks in the company has more skin in the company and will want to remain in the company to ensure that the company performs well (Gong et al., 2017). Every unit increase in *YearsSinceLastPromotion* is associated with a 0.23 decrease in the log-odds of attrition, all else equals. This seems counterintuitive because it is expected that a lack of career progression will impel employees to resign (Morris, 2018). A possible explanation for the observation of the opposite effect is that the company operates in a niche sector in healthcare and there is a lack of exit opportunities, an effect that will become further entrenched the longer an employee works at the company (Zacher et al, 2015). A secondary reasoning is parallel to the explanation for the estimated coefficient for *PercentSalaryHike* – an employee may time their resignation to be right after a promotion in order to gain leverage in salary negotiation. Every unit increase in *BusinessTravel* is associated with 0.65 increase in log-odds probability. Several explanations may account for such an observation. Employees may dislike excessive business travel as it may be exhausting in the long run (Burns, 2019). Selection effect could also account for this observation, where high competency individuals are both those who have the opportunity to travel and network frequently and are also more often poached by competitors.

Part II: Assessment of company-related factors affecting attrition of new young-to-mid-career company hires

To gain a more nuanced understanding of the associations between the aforementioned factors and *Attrition*, I investigated the heterogenous treatment effects of the predictor variables on *Attrition* by examining the subpopulation coefficient estimates of the hierarchical models. Understanding variations in the magnitude of associations across subpopulations allow the company to understand how company policies affect different age groups and job roles, and implement targeted policies to maximise the effectiveness of talent retention. Importantly, examining the variations in effects across subpopulations may also provide insights into the competing explanations behind the effects of *YearsSinceLastPromotion* and *BusinessTravel*, as expounded upon above.

Figure 5 consists of the 95% credible intervals for the coefficient estimates of each *AgeGroup*. The coefficient estimates for *PercentSalaryHike* and *StockOptionLevel* are largely homogenous across age groups. The effect of *YearsSinceLastPromotion* on *Attrition* varies substantially across age groups and are mostly negative except for employees between 22-23 years of age, for which the coefficient estimate is positive. This anomaly can be explained by the fact that 22- and 23-year-olds, who have a long career horizon ahead of them, are unwilling to invest their career in a company that does not reward their contribution with a commensurate promotion. However, this explanation may not be well-founded given that other employees in their early 20s do not necessarily exhibit a tendency to attrit when they are not promoted. Furthermore, the 95% credible interval of the coefficient estimate contains zero, suggesting that the effect of a lack of promotion on 22- and 23-year-olds may in fact be negligible.

Figure 5: Credible intervals of parameter estimates of hierarchical model (age group)



The effects of *YearsSinceLastPromotion* and *BusinessTravel* on *Attrition* also vary substantially across age groups. Interestingly, the intensity of the effect of *BusinessTravel* seem to decline with age, which suggests that younger employees with more opportunities for *BusinessTravel* are more likely to leave the company than older employees. Relating back to the competing explanations on why more frequent *BusinessTravel* culminates in higher *Attrition*, this observation seems to be in favour of the case of the aforementioned "selection effect" over the "exhaustion effect". Should the latter be the dominant reason for attrition, one should observe a positive correlation between age and the intensity of the effect since younger employees, in general, are expected to be more excited for exposure on business trips. In contrast, a negative correlation between age and the intensity of the *BusinessTravel* effect is in tandem with the explanation of the "selection effect", since young talents are likely to have greater growth potential and are more highly sought after compared to older talents.

Figure 6: Credible intervals of parameter estimates of hierarchical model (job role)

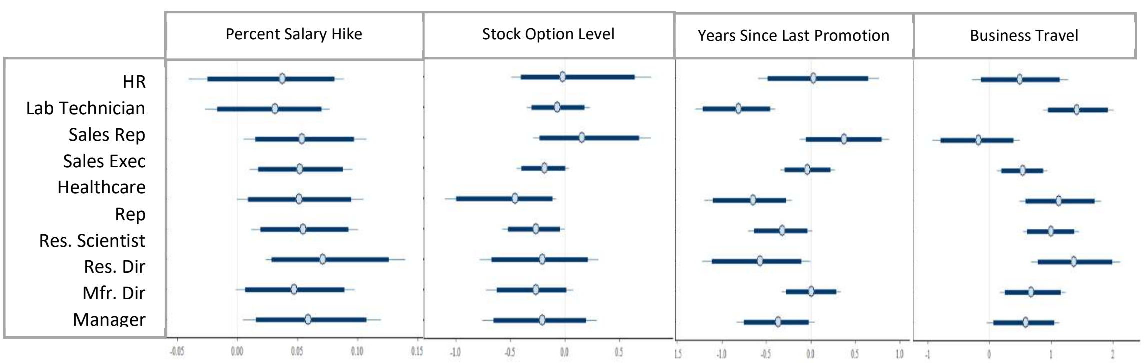


Figure 6 reports the 95% credible intervals of coefficient estimates for all four predictor variables across job roles. Consistent with the standard error estimates, effects of *PercentSalaryHike* and *StockOptionLevel* are largely homogenous across job roles. For majority of the technical job roles (i.e. laboratory technician, healthcare representative, research scientist, and research director), *YearsSinceLastPromotion* is negatively associated with *Attrition*. On the other hand, for generic job roles (i.e. human resources, sales representative and sales executive), *YearsSinceLastPromotion* has weaker to no negative effect on *Attrition*, and there is even a positive association between *YearsSinceLastPromotion* and *Attrition* for sales representatives. The larger magnitude of the effects of *YearsSinceLastPromotion* on technical compared to generic job roles supports the explanation of the "entrenchment effect", where the limited exit opportunities due to the nicheness of the company's operations are likely to affect employees in technical jobs more than generic jobs. In contrast, the “bargaining power” explanation should entail a relatively homogenous effect across job roles, which we do not observe here. Interestingly, sales representatives also appear to be least affected by *BusinessTravel.* In line with the reasoning of the proposed “selection effect”, this could be because going on frequent business trips is a job requirement of sales representatives. As opposed to other job roles where business trips mainly accrue to high-performing individuals, there is therefore limited selection effect of high-performance sales representatives.

4. Conclusion and recommendations

Attrition rate in a company one of the foremost concerns of a HR function due to the associated costs in knowledge loss, project disruption and recruitment. In this project, I modelled the attrition tendencies of employees in an anonymised healthcare company with an above-average attrition rate using Bayesian logistic regression models and hierarchical models. By comparing the models based on model complexity and model fit, this paper found that the ridge regression model has the best performance in modelling the effects of *PercentSalaryHike, StockOptionLevel, YearsSinceLastPromotion* and *BusinessTravel* on *Attrition*. This insight is valuable for the company as it allows them to effectively predict and pre-empt imminent attritions and helps to streamline manpower requirements, improving operational efficiency and reducing costs.

This paper also investigated the associations between various company-related variables and the attrition tendencies of employees. The positive association between *PercentSalaryHike* and *Attrition* is likely to reflect employees’ wont of negotiating for a salary raise before attrit in order to augment their bargaining power when joining a new company. *StockOptionLevel* is negatively associated with *Attrition*, which is likely to be because higher *StockOptionLevel* incentivises employees to stay in the company to ensure the company’s good performance. *YearsSinceLastPromotion* is negatively associated with *Attrition*, with a stronger association for technical job roles. This paper proposes the “entrenchment effect”, where the specificity of the company’s work limits exit opportunities of employees. Lastly, *BusinessTravel* is positively associated with *Attrition*, with a stronger association for young employees. This can be accounted for by the “selection effect”, where high-performance employees, who tend to be exposed to more *BusinessTravel*, are also more likely to find job opportunities with other companies.

These insights are valuable for augmenting the efficacy of talent retention strategies. To improve talent retention of young-to-mid-career company hires, this paper recommends offering higher *StockOptionLevel(s)* for high-performing individuals. Importantly, this paper distinguished higher *StockOptionLevel(s)* as an effective incentive mechanism relative to other potential incentives such as *PercentageSalaryHike*, *MonthlyIncome*, and *YearsSinceLastPromotion*. The fact that these factors do not exhibit a negative association with *Attrition* also seems to indicate that employees are generally satisfied with the salary, salary increment and promotion rate that they are currently receiving. As offering higher *StockOptionLevel(s)* may be costly to the company, the company should improve the efficacy of its talent retention programme by focusing its attention on the needs of employees who are less than or equal to 30 years old with frequent *BusinessTravel* exposure, as these are employees with the highest tendency to attrit. This paper recommends conducting a detailed company satisfaction survey with this group of employees to further narrow down areas of improvement in the company’s HR policies.

Bibliography

Booz, M. (2018). These 3 Industries Have the Highest Talent Turnover Rates. *Linkedin.* Retrieved from: https://business.linkedin.com/talent-solutions/blog/trends-and-research/2018/the-3-industries-with-the-highest-turnover-rates

Brownlee, J. (2020). Probabilistic Model Selection with AIC, BIC, and MDL. *Machine Learning Mastery*. Retrieved from: https://machinelearningmastery.com/probabilistic-model-selection-measures/

Burns, K. (2019). Is Frustrating Corporate Travel Leading to Employee Turnover? *Lola.* Retrieved from: https://www.lola.com/blog/is-frustrating-corporate-travel-sending-your-best-employees-packing

Gelman et al. (2008). A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models. *The Annals of Applied Statistics, 2008, Vol. 2, No. 4*. Retrieved from: http://www.stat.columbia.edu/~gelman/research/published/priors11.pdf

Gong et al. (2017). Retention Effects of Employee Stock Options: Evidence from Bunching at Vestings Dates. *University of Pennsylvania.* Retrieved from: https://economics.sas.upenn.edu/system/files/2018-03/draft20171014_0.pdf

McLaren, S. (2019). Here's How IBM Predicts 95% of its Turnover Using Data. *Linkedin.* Retrived from: https://business.linkedin.com/talent-solutions/blog/artificial-intelligence/2019/IBM-predicts-95-percent-of-turnover-using-AI-and-data

Morris, S. (2018). Lack of Career Development Drives Employee Attrition. *Smarter with Gartner.* Retrieved from: https://www.gartner.com/smarterwithgartner/lack-of-career-development-drives-employee-attrition/

O'Connor, S. W. (2020). Human Resources Analytics: What It Is and Why It's Important. *Northeastern University.* Retrieved from: https://www.northeastern.edu/graduate/blog/human-resources-analytics/

Singh, A. (2020). Exploratory Data Analysis – Employee Attrition Rate. *Medium.* Retrieved from: https://medium.com/swlh/exploratory-data-analysis-employee-attrition-rate-591ce8e7518f#:~:text=For%20a%2095%25%20confidence%20interval,between%2014.3%25%20and%2018%25.

Swathi, S. (2020). A View on Employee Attrition. *International Journal of Scientific and Research Publications, Volume 10, Issue 3.* Retrieved from: http://www.ijsrp.org/research-paper-0320/ijsrp-p9905.pdf

Tiwari et al. (2007). Selecting the Appropriate Outlier Treatment for Common Industry Applications. *NESUG Statistics and Data Analysis.* Retrieved from: https://www.lexjansen.com/nesug/nesug07/sa/sa16.pdf

Van Erp et al. (2019). Shrinkage Priors for Bayesian Penalised Regression. *Journal of Mathematical Psychology.* Retrieved from: https://osf.io/cg8fq/download

Zacher et al. (2015). Career Adaptibility and Career Entrenchment. *Journal of Vocational Behavior, 88(1), 164-173.* Retrieved from: https://www.researchgate.net/publication/274095690_Career_Adaptability_and_Career_Entrenchment