Globex Pharma Employee Attrition
A Predictive Analysis

Name: Amanda Wright

Due Date: 17th June 2024

Executive Summary

Globex Pharma, focused on better understanding the reasons behind employee turnover, has requested a statistical analysis of the possible reasons for attrition in their ranks to reduce the costs to the organisation,

The study uses machine learning algorithms in Logistic Regression and Decision Trees in conjunction with the Globex Pharma Employee Survey data to produce two models that give us insight into the likely characteristics of an employee that has left Globex since the 2023 survey.

Using the data the models have indicated there is a link between overtime performed by an employee, the number of years an employee has been in the workforce being under 3 years, marital status being single and being a frequent business traveller to name a few.

Further to those factors, there are high income earners particularly in the Sales departments that have no stock options which additionally correlated to higher levels of turnover in that group.

We have provided recommendations based on the models to reduce overtime, develop a strategy to retain younger employees, review all stock option allocations for consistency in the organisation, perform independent exit interviews moving forward and a review of the travel needs for employees, particularly in the sales department.

A focus on these areas should lead to improved attrition and reduction in the economic impacts for Globex Pharma.

An additional review in 6-12 months should be undertaken to review the effects of the efforts laid out in this document for efficacy.

Table of Contents

Introduction

Since the Globex Employee Survey in 2023 16.2% of Globex's one thousand employees have left the organisation resulting in lost sales and increased costs in recruiting, induction and onboarding replacements with the total cost difficult to quantify.

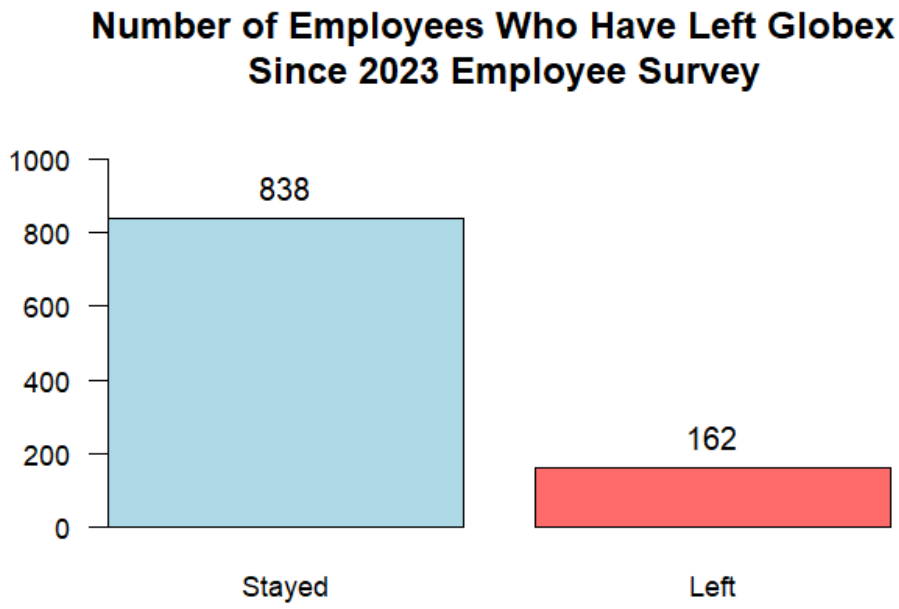


Fig. 1 Employee Attrition Globex Employee Survey 2023

With a view to reduce the economic impact of attrition moving forward, we will complete a statistical analysis resulting in recommendations aimed at reducing attrition in the organisation, and in turn reduce the costs involved.

Utilising the Globex Pharma Employee Survey 2023 in combination with new data released to us on employee turnover, two separate predictive models will be created.

Each model, using separate statistical methods, will analyse the historical data on who has left previously, to provide insights on who may will leave and stay at Globex in the future.

Each model, using machine learning algorithms, will generate a set of *characteristics of employees* that are most likely to be present when an employee leaves, to a certain level of accuracy.

We will use the characteristics most likely to be present in the "Leavers", to make recommendations on what actions may best reduce the cost of attrition at Globex moving forward.

The goal is to answer the question:

"What attributes of an employee *increase the likelihood* of leaving Globex Pharma?"

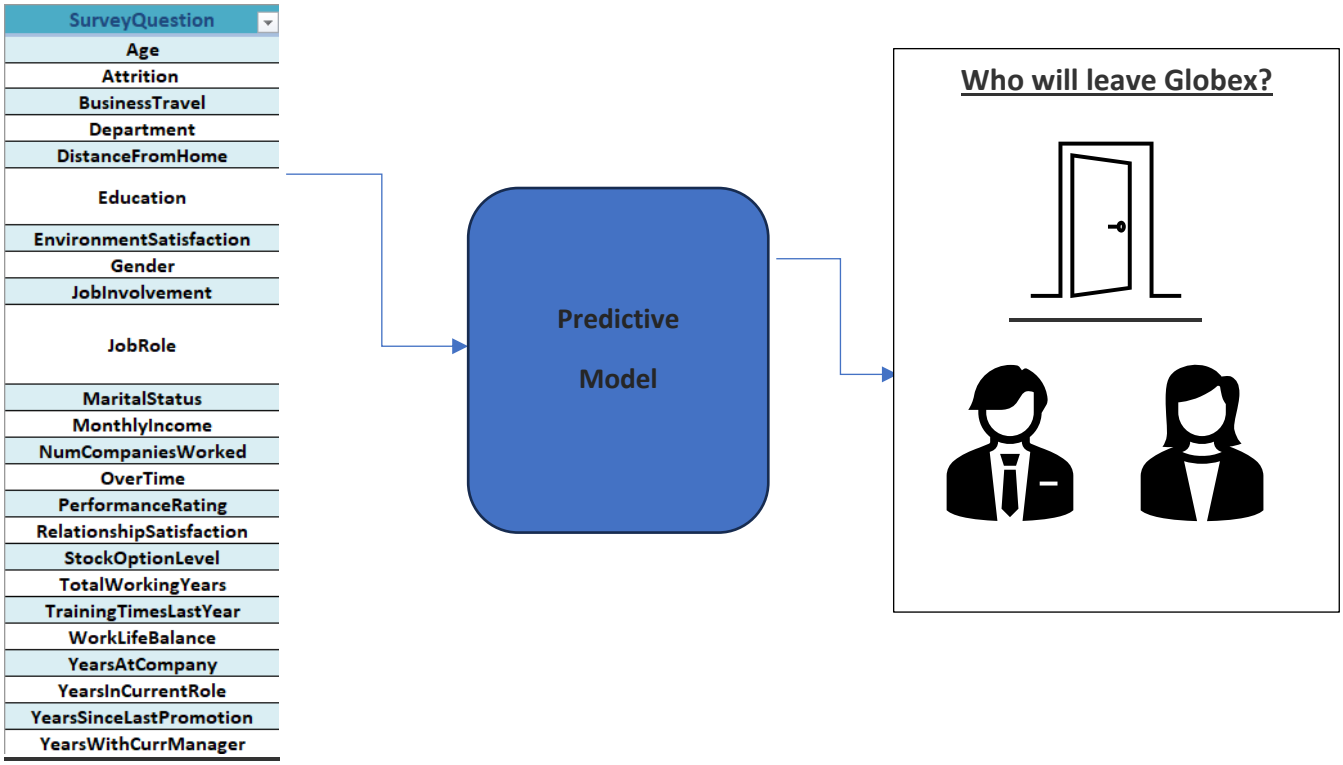


Fig. 2 Predictive Analytics Process, Refer to Appendix A for Globex Survey Questionnaire details

Analysis

Overall Methodology

The first model will follow the methodology of Logistic Regression, the second of Decision Trees. (Refer to Appendix B Glossary)

Logistic Regression is a supervised machine learning algorithm utilised in predicting outcomes of a binary nature, such as yes/no or true/false, from a set of predictor variables. ("Logistic regression | Definition & Facts | Britannica," 2024)

Decision Trees use a flowchart like structure to make decisions towards a prediction, like branches in a tree or a fork in the road, used to predict the likelihood of an outcome occurring such as an employee has left or not left. ("Decision Tree," 2017)

For both methods, we have taken the twenty-three questions from the survey and using the popular R programming language to build a series of models based on different combinations of the survey question.

The survey questions are called "predictor variables" in that we will use them to predict the target variable "Attrition" – whether an employee will stay or leave.

Next step is, for each unique combination of predictor variables we will compare the accuracy of the model's predictions.

We can use this comparison to decipher which model was most accurate based on the historical data, or which combination of survey questions/predictor variables did the "best" job at predicting who stayed and left Globex in the past, to gain insights into what may happen in the future.

Predictive Model One – Logistic Regression

Predictive Model One - "What attributes make an employee *more likely* to leave Globex Pharma?"

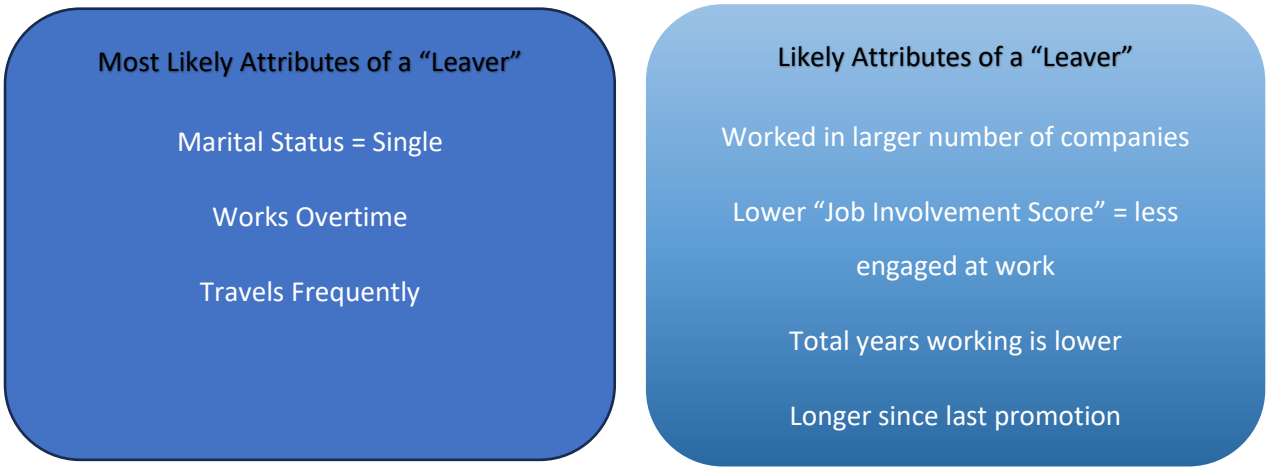


Fig. 3 Logistic Regression Model – factors leading to increased rates of Attrition at Globex

The key characteristics likely to be in our "Leavers" from the Logistic Regression model are stated in Fig 3.

We feel this model gives Globex the best chance to focus efforts on improving Attrition.

From our analysis (using the assumptions in Appendix B) with a goal to gain the best model with between five and eight variables, this was the chosen model.

Fig 3 lays out the model output in terms of the "Most Likely Attributes of a Leaver" and a group of "Likely Attributes of a Leaver" that are still statistically significant but less associated then the first set with an employee predicted to leave Globex.

Accuracy of the Logistic Regression Model

The "best" model in our analysis has twelve predictor variables which is overly-complex with limited effective use in this scenario, and additionally difficult to measure the effect of any effort to impact attrition.

Hence we have chosen a simpler model with six variables with similar accuracy, albeit slightly lower. Refer Appendix C for detailed model outputs and breakdown of the modelling process.

Assuming that the environment at Globex is not too different from when the survey was completed in 2023, and assuming the survey is accurate data to begin with we can use the attributes from the model to focus our recommendations on with confidence.

If there have been disruptive changes in the organisation or the survey data is poor, the predictions of the model could therefore be less useful in determining attrition levels and impacting them positively.

Predictive Model Two – Decision Tree

Predictive Model Two - "What attributes make an employee *more likely* to leave Globex Pharma?"

Given the assumptions in the methodology given in Appendix B, the output generated by Predictive Model Two is the Decision Tree in Fig 5 and summarised in Fig 4.

The "leaves" of the tree that end in a "green ellipse" in Fig 5 are where the model predicts the employee will leave, blue means employees are predicted to stay.

The probability of employees in that bucket leaving is between 0 and 1 and given in each box both green and blue.

Fig 4 estimates the financial cost to replace an employee if the predictions are accurate.

According to Forbes magazine (Bergstrom, n.d.) the cost for a highly skilled employee is up to 213% of an average employee wage.

The median wage at Globex is $4864, however the median wage for employees less than 3 years is $2372. For the sake of the exercise, we will estimate the cost of attrition at $5000 per employee.

Attributes	Probability of Leaving	% of Workforce	Number of Employees	Cost of Attrition $5000 per person
Total Working Years >=3, Works Overtime, Income >= $3924, No Stock Options, Department = Sales, Income >= 8164	80%	1%	10	80% x 10 x $5000 = $40K
Total Working Years < 3 years, Works Overtime	71%	3%	28	71% x 28 x $5000 = $99.4K
Total Working Years >= 3 years, Works Overtime, Salary < $3924, Training Times >= 4	68%	2%	19	68% x 19 x $5000 = $64.6K
Total Working Years < 3 years, No Overtime, Department = Sales	67%	2%	18	67% x 18 x $5000 = $60.3K
TOTAL	**-**	**7.5%**	**75**	**$264,300**

Fig. 4 Summary of the Decision Tree in Fig. 5

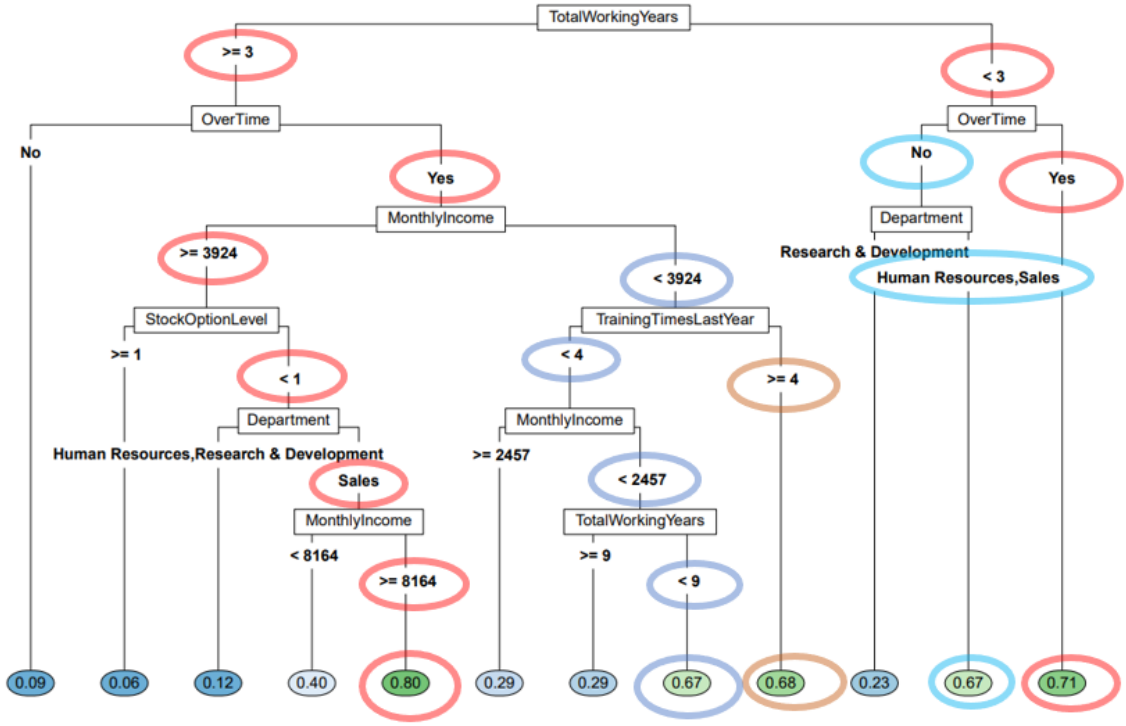


Fig. 5 Decision Tree Model Output – Probability (0 to 1) of leaving in the "leaves", Employees predicted to leave in green "leaves" and circled

Accuracy of the Decision Tree Model

The "best" model with the highest accuracy found by our analysis has eleven predictor variables again an overly complex model with reduced usability, therefore we have balanced a less accurate model with a simpler model.

We have therefore chosen the "best" model with six predictor variables, with slightly less accuracy. (Refer Appendix D for detailed model outputs)

A point to note, this model is less accurate compared to the logistic regression model.

Conclusion

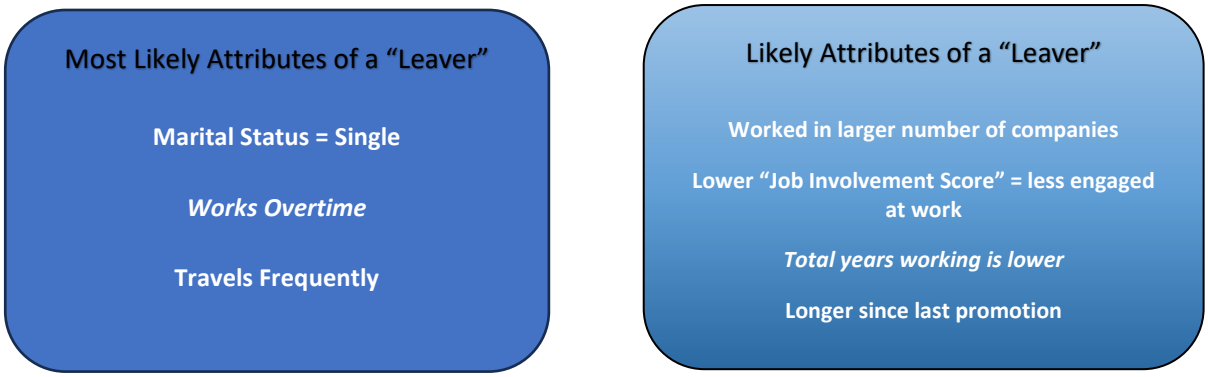
We conclude that we have been successful in finding factors that seem to have been present in the employees that have left Globex since the Employee Survey in 2023, with a reasonable level of accuracy and predictability for the future in our models.

Interpretations of Results

- Whether an employee works "Overtime" is important, and strongly present in both models
- "Total Working Years" is a clear commonality, and the Decision Tree highlights less than 3 years in the workforce may increase the probability of attrition significantly.

- "Total Working Years" firstly appears in the Logistic Regression model as a factor, and potentially secondly in the "Highly Likely" attributes as "Marital Status = Single" which is closely related to "Total Working Years" < 3 years.
- Being in the "Sales" department appears in the Decision Tree twice and possibly in Logistic Regression model as "Business Travel – Frequently" and low "Job Involvement" rating as employees who travel frequently are difficult to keep engaged. ("Managing Sales Rep Turnover," n.d.)
- "Stock Option Level" and "Monthly Income" feature in the high-income earners in the Sales department with no stock options leading to a higher probability of leaving

Logistic Regression Model



Decision Tree Model

Fig. 6 Predictive Model Outputs – Characteristics of the "Leavers"

Recommendations

Recommendation One: An Internal Review of Overtime

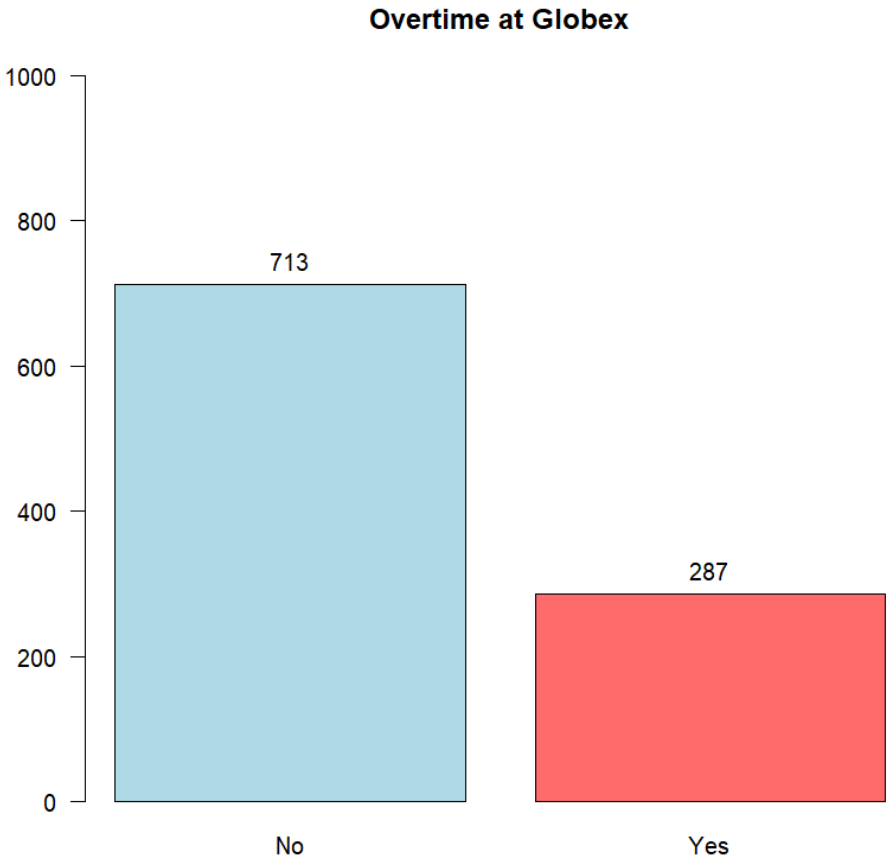


Fig. 7 Overtime at Globex – Employee Survey 2023

According to Ko and Choi's (2019, p. 284) study into time lagged data of 273 firms, and Harvard Business Review ("It's Time to Prioritize Your Work-Life Balance | Harvard Online," n.d.) the research is clear overtime is linked to:

- Deterioration in physiological and psychological health
- Is negatively related to personal satisfaction at work, whilst positively related to firm productivity
- Increased rates of depression and anxiety
- Increased negativity between cohorts which can affect cultures quickly amongst sub organizations

With almost 30% of employees at Globex completing overtime, and its prevalence in our models as a factor for attrition we recommend a full review of overtime at Globex led by the Finance Team with a view to heavily decrease and potentially remove overtime at Globex, without losing firm productivity gains.

Our research suggests the cost financially of overtime does not make great business sense as the work produced in overtime hours is generally poorer quality and lower volume due to fatigue. ("This is how shorter working hours can affect your productivity," 2021)

Questions to answer in the review are:

- Who is completing overtime and why?
- What is the total cost in $ of this overtime, can that money be utilised more efficiently elsewhere?
- When is overtime being completed? In a specific period of the year, a busy season?
- What mechanisms are in place to track overtime regularly?
- Do employees feel the need to work overtime to supplement their salary?

Particularly focus on junior employees less than 3 years in the workforce, as both models suggests could be a strong correlation to attrition, and research literature suggests that Gen Z and Millennials are more focused on work-life balance than older generations. (Kelly, n.d.)

Recommendation Two : Strategize Engagement of Gen Z and Millennials

Both models suggest attrition rates are high in this group

With 8% of the workforce at Globex total working years less than 3 years (Fig. 8), risk of cost here is evident.

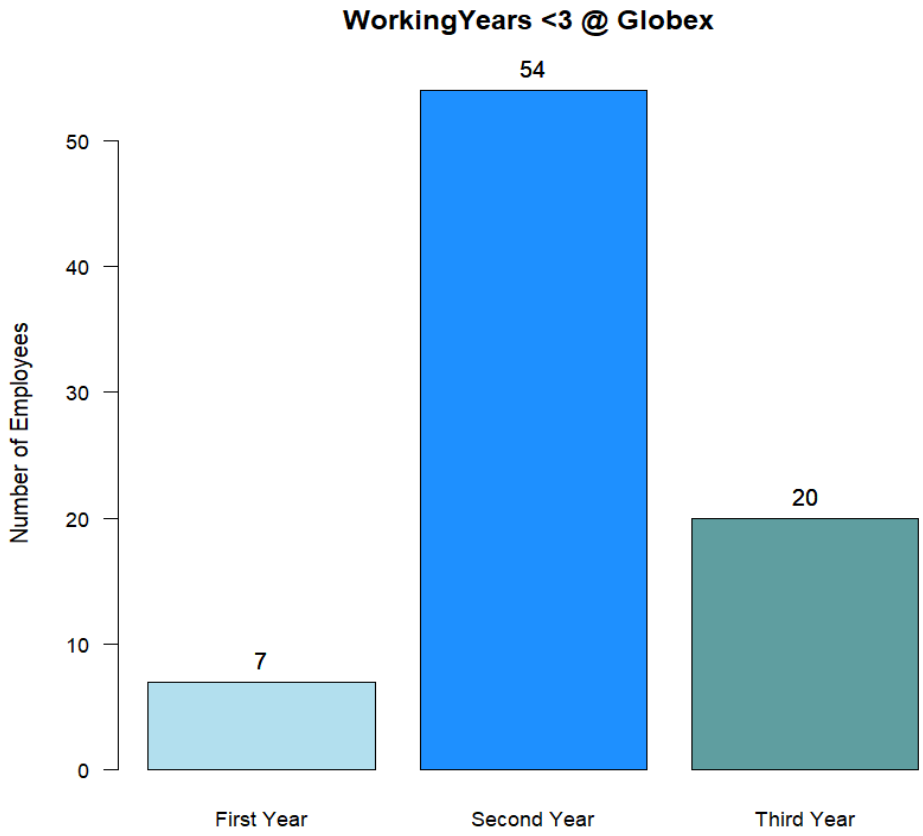


Fig. 8 Employees Less Than 3 Years – Survey Data

Global trends show attrition will continue to be high in this category as research suggests on average 91% of Millennials plan to spend less than 3 years with any employer. (“Boost Global Workforce Retention By Overcoming These Five Challenges,” n.d.)

Therefore, the priority here is to create a strong strategy to minimise the loss of junior employees, particularly those with high potential, in the early years, remembering the cost of replacing an employee is conservatively estimated at $5000.

We recommend Globex create a strategy document with Human Resources and Senior Execs considering:

- a strong mentoring plan and pathway for high potential junior employees, which can lead to improve retention of employees (Sahai, n.d.)
- perks and benefits focused more on mental health such as gym, yoga and travel (“The Benefits of Corporate Fitness Programs in Reducing Employee Turnover,” 2024)
- flexible work schedules or reduced working hours are the preference of up to 75% of this group (Kelly, n.d.)
- overtime should be minimised and not considered the “norm” to get ahead (Kelly, n.d.), given work-life balance is a priority to this group
- nail the engagement from “Day One” - review onboarding policies and procedures of first year employees (Tenakwah, 2021)
- Follow up with regular anonymous surveys with employees in first three years of work, every 6-12 months

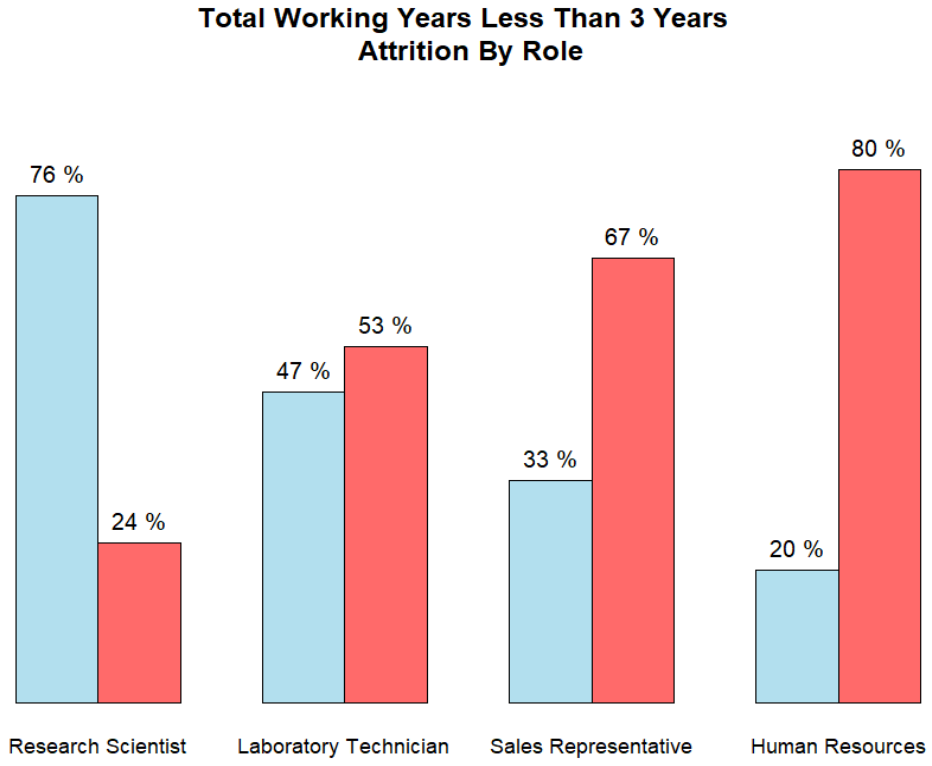


Fig. 9 Employees Less Than 3 Years By

Recommendation Three : Renew Business Travel Criteria

- Review of criteria for business travel, ensuring it is cost effective with a return on investment, and the process for approval of travel needs.
- Where travel is deemed necessary and approved, consider a clear plan for engagement with team members in travelling roles.
- Research here suggests a consistent catch-up process with team on the road to ensure they are feeling a sense of belonging and involved heavily in the organisation. ("Managing Sales Rep Turnover," n.d.)

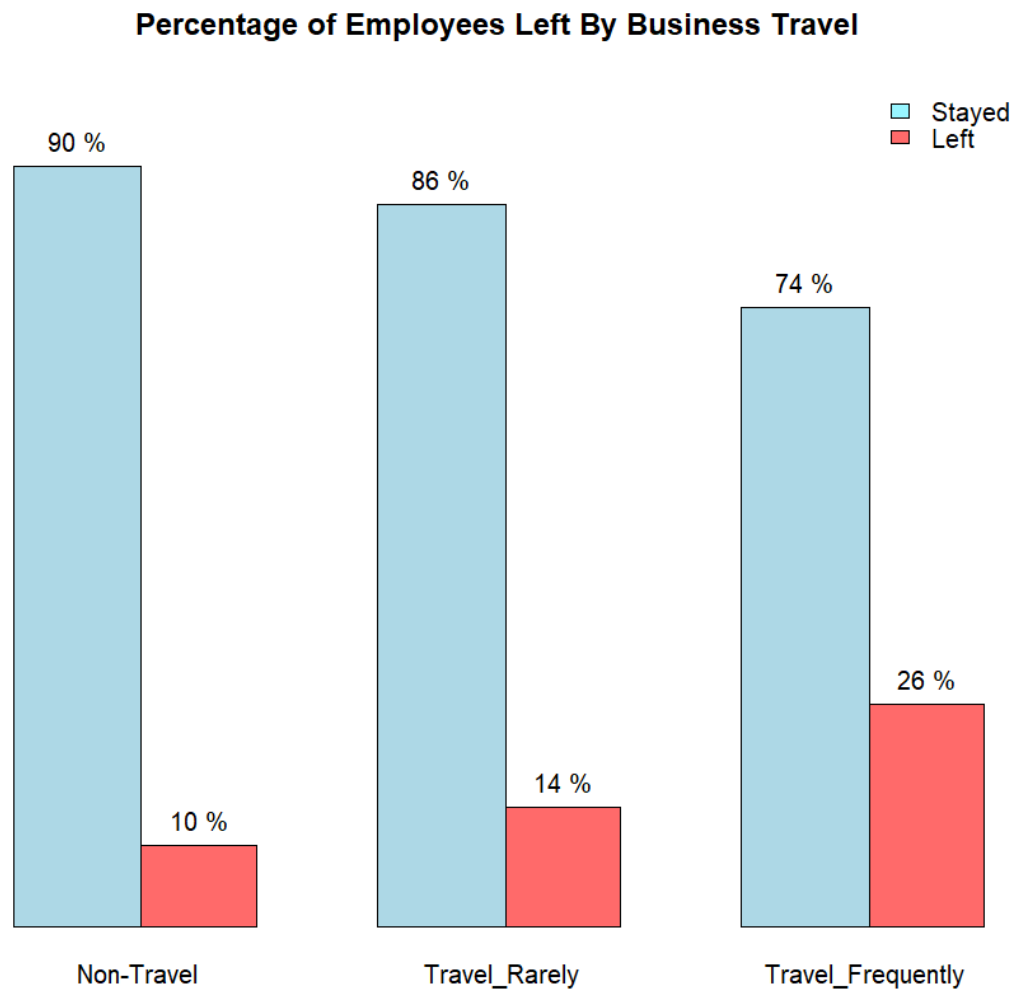


Fig. 10 Business Travel Attrition Rate

Recommendation Four : Review Stock Option Policy

- Review Stock Option allocation to ensure consistency particularly those on high salaries, and across departments.

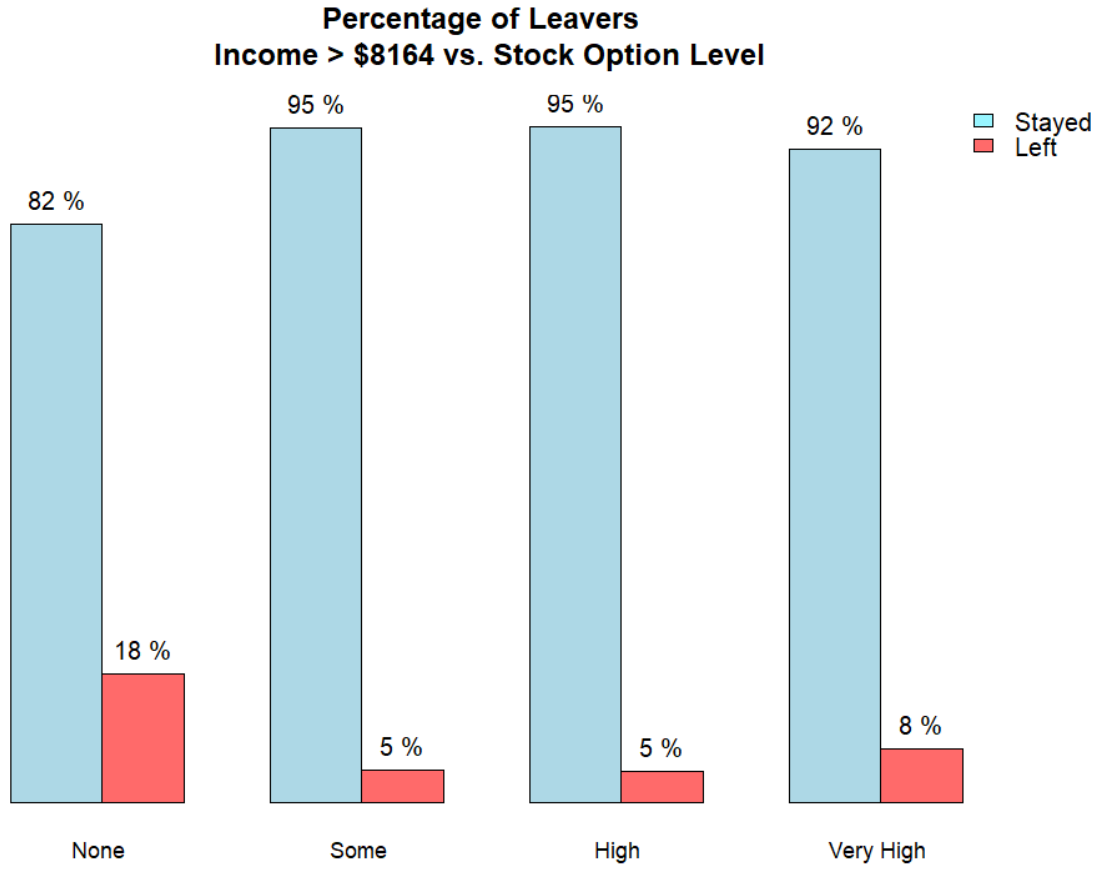


Fig. 11 High Income Earners Attrition By Stock Option Level

Recommendation Five: Independent Exit Interviews

- Engage a consultant for recommendations to complete an independent collection of exit interviews for all employees, via email link when terminated. (Spain and Groysberg, 2016)
- Review every 12 months

References

Bergstrom, M., n.d. Council Post: Why Employee Churn Is Killing Your Company [WWW Document]. Forbes. URL https://www.forbes.com/sites/forbestechcouncil/2022/08/16/why-employee-churn-is-killing-your-company/ (accessed 6.11.24).

Boost Global Workforce Retention By Overcoming These Five Challenges [WWW Document], n.d. . Velocity Global. URL https://velocityglobal.com/resources/blog/increase-global-workforce-retention/ (accessed 6.13.24).

Decision Tree [WWW Document], 2017. . GeeksforGeeks. URL https://www.geeksforgeeks.org/decision-tree/ (accessed 6.5.24).

How to interpret AUC score (simply explained) [WWW Document], 2022. . Stephen Allwright. URL https://stephenallwright.com/interpret-auc-score/ (accessed 6.17.24).

Kelly, J., n.d. Gen-Zers And Millennials Are Opting Out Of The Traditional Corporate Climb [WWW Document]. URL https://www.forbes.com/sites/jackkelly/2024/06/07/gen-zs-and-millennials-are-opting-out-of-the-traditional-corporate-climb/ (accessed 6.12.24).

Ko, Y.J., Choi, J.N., 2019. Overtime work as the antecedent of employee satisfaction, firm productivity, and innovation. Journal of Organizational Behavior 40, 282–295. https://doi.org/10.1002/job.2328

It's Time to Prioritize Your Work-Life Balance | Harvard Online [WWW Document], n.d. URL https://www.harvardonline.harvard.edu/blog/prioritize-work-life-balance (accessed 6.13.24).

Logistic regression | Definition & Facts | Britannica [WWW Document], 2024. URL https://www.britannica.com/science/logistic-regression (accessed 6.5.24).

Managing Sales Rep Turnover [WWW Document], n.d. URL https://www.linkedin.com/pulse/managing-sales-rep-turnover-john-rankins-g1i9c (accessed 6.13.24).

Sahai, K., n.d. Council Post: Reduce Turnover With A Culture Of Coaching [WWW Document]. Forbes. URL https://www.forbes.com/sites/forbescoachescouncil/2018/01/29/reduce-turnover-with-a-culture-of-coaching/ (accessed 6.13.24).

Spain, E., Groysberg, B., 2016. Making Exit Interviews Count. Harvard Business Review. https://hbr.org/2016/04/making-exit-interviews-count (accessed 6.13.24)

Statistical Significance: Definition, Types, and How It's Calculated [WWW Document], n.d. . Investopedia. URL https://www.investopedia.com/terms/s/statistical-significance.asp (accessed 6.17.24).

Tenakwah, E.S., 2021. What do employees want? Halting record-setting turnovers globally. Strategic HR Review 20, 206–210. https://doi.org/10.1108/SHR-08-2021-0040

The Benefits of Corporate Fitness Programs in Reducing Employee Turnover [WWW Document], 2024. . Employee Wellness Australia. URL https://employeewellnessaustralia.com.au/the-benefits-of-corporate-fitness-programs-in-reducing-employee-turnover/ (accessed 6.13.24).

This is how shorter working hours can affect your productivity [WWW Document], 2021. . World Economic Forum. URL https://www.weforum.org/agenda/2021/09/teams-productivity-work-hours/ (accessed 6.13.24).

Appendices

Appendix A - Globex Survey Questions 2023

SurveyQuestion	QuestionDetail
Age	Age in years
Attrition	Whether the employee left during the last year (Yes/No)
BusinessTravel	Business travel frequency (Non-Travel, Travel_Frequently, Travel_Rarely)
Department	Department (Human Resources, Research & Development, Sales)
DistanceFromHome	Number of kms from home to work
Education	Highest level of education: 1=High School, 2=Diploma, 3=Bachelor, 4=Masters, 5=PhD
EnvironmentSatisfaction	Satisfaction with work environment: 1=Low, 2=Medium, 3=High, 4=Very high
Gender	Male/Female
JobInvolvement	1=Not engaged, 2=Somewhat not engaged, 3=Somewhat engaged, 4=Engaged
JobRole	Healthcare Representative, Human Resources, Laboratory Technician, Manager, Manufacturing Director, Research Director, Research Scientist, Sales Executive, Sales Representative
MaritalStatus	Divorced/Married/Single
MonthlyIncome	Monthly net pay ($)
NumCompaniesWorked	Number of other companies worked for
OverTime	Whether normally work overtime (Yes/No)
PerformanceRating	1=Low, 2=Medium, 3=High, 4=Very high
RelationshipSatisfaction	Satisfaction with work colleagues: 1=Low, 2=Medium, 3=High, 4=Very high
StockOptionLevel	Stock option levels in company: 0=None, 1=Some, 2=High, 3=Very high
TotalWorkingYears	Number of years at work (all jobs)
TrainingTimesLastYear	Training days in last year
WorkLifeBalance	1=Bad, 2=Satisfactory, 3=Good, 4=Very good
YearsAtCompany	Number of years at company
YearsInCurrentRole	Number of years in current role
YearsSinceLastPromotion	Number of years since last promotion
YearsWithCurrManager	Number of years with current manager

Appendix B – Methodology and Assumptions

Given the twenty-three variables in the survey there are $2^{23} \approx 8.4\ million$ combinations of the predictor variables which can be tested to find the most accurate model, however even in the 21st Century the task of testing them all is difficult in a desired timeframe.

We have used the assumptions below to reduce the combinations for the models to assess:

- Our goal is to have between five and eight variables – a simpler model means less areas to focus to reduce attrition. The number of variables is an arbitrary choice we have made here for simplicity and has no statistical relevance.
- Given this assumption, we will be looking for the "best" model that can be found between five and eight variables, less than five may oversimplify the factors that are causing employees to leave Globex, resulting in less effective recommendations.
- "Best" will be defined here by our choice of using the ten-fold cross validation method and the highest AUC(Area Under the Curve) measure between 0 and 1, (Refer to Glossary for more information)

- Using Backward Subset Selection(Refer to Glossary, Appendix C and D), where the "best" AUC score results in the model being overly complex we will accept a slightly lower AUC score to get a simpler model.

Glossary

AUC Score	AUC is a common abbreviation for Area Under the Receiver Operating Characteristic Curve (ROC AUC). It is a metric used to assess the performance of classification machine learning models. ("How to interpret AUC score (simply explained)," 2022) **AUC score** — **Interpretation** >0.8 — Very good performance 0.7-0.8 — Good performance 0.5-0.7 — OK performance 0.5 — As good as random choice
Best Subset Selection	Best model (highest AUC score) of all combinations of variables, in this case $2^{23} \approx 8.4\ million$
Backwards Stepwise Selection	When selecting variables in a model, starts with all variables included and then the least significant one taken out one by one, compare with Forward Subset Selection which starts with no variable and adds the most significant one each time.
Decision Trees	Decision Trees use a flowchart like structure to make decisions towards a prediction used to predict the likelihood of an outcome occurring such as an employee has left or not left. ("Decision Tree," 2017)
Logistic Regression	Logistic Regression is a supervised machine learning algorithm utilised in predicting outcomes of a binary nature, such as yes/no or true/false, from a set of predictor variables. ("Logistic regression | Definition & Facts | Britannica," 2024)
Statistical Significance	Statistical significance refers to the claim that a set of observed data are not the result of chance but can instead be attributed to a specific cause. ("Statistical Significance," n.d.)

Appendix C – Predictive Model Logistic Regression One Outputs

"Best Model"

Accuracy - AUC Score 0.8063
Number of Predictors - 12

Method

Logistic Regression using Ten-Fold Cross Validation and Backward Stepwise Selection

Backward Subset Selection was chosen with a view it has the best chance of the most important variables being in the final model, and in our trials have proven to have a better AUC score and more significant predictor variables as outputs in the models we have run.

```
Best predictors: BusinessTravel + DistanceFromHome + EnvironmentSatisfaction + JobInvolvement + MaritalSta
tus + NumCompaniesWorked + OverTime + RelationshipSatisfaction + StockOptionLevel + TotalWorkingYears + Wo
rkLifeBalance + YearsSinceLastPromotion
> message("Cross-validation score: ", cv_score(best_predictors))
Cross-validation score: 0.806394764160787
>
> # Build the final model from the predictors
> final_model <- glm(
+
+   formula = paste("Attrition ~ ", paste(best_predictors, collapse = " + "), sep = ""),
+   data = employees,
+   family = "binomial"
+ )
>
> summary(final_model)

Call:
glm(formula = paste("Attrition ~ ", paste(best_predictors, collapse = " + "),
    sep = ""), family = "binomial", data = employees)

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                    0.56019    0.80342   0.697 0.485643
BusinessTravelTravel_Frequently  1.61016    0.44838   3.591 0.000329 ***
BusinessTravelTravel_Rarely      0.69629    0.41806   1.666 0.095812 .
DistanceFromHome                 0.02896    0.01218   2.377 0.017450 *
EnvironmentSatisfaction         -0.36257    0.09290  -3.903 9.50e-05 ***
JobInvolvement                  -0.57916    0.13972  -4.145 3.40e-05 ***
MaritalStatusMarried             0.46096    0.32234   1.430 0.152707
MaritalStatusSingle              1.37941    0.40533   3.403 0.000666 ***
NumCompaniesWorked               0.17793    0.04065   4.377 1.20e-05 ***
OverTimeYes                      1.61598    0.20861   7.746 9.46e-15 ***
RelationshipSatisfaction        -0.26549    0.09034  -2.939 0.003296 **
StockOptionLevel                -0.16681    0.17941  -0.930 0.352482
TotalWorkingYears               -0.14383    0.02128  -6.757 1.40e-11 ***
WorkLifeBalance                 -0.30586    0.13472  -2.270 0.023184 *
YearsSinceLastPromotion          0.16564    0.03974   4.168 3.08e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 885.94  on 999  degrees of freedom
Residual deviance: 669.57  on 985  degrees of freedom
AIC: 699.57

Number of Fisher Scoring iterations: 6
```

"Chosen Model"

Accuracy - AUC Score 0.7880, accepting a 0.007 reduction in AUC for simpler model
Number of Predictors - 7
Method - Logistic Regression using Ten-Fold Cross Validation Backward Stepwise Selection

Note: We have removed “Business Travel - Rarely” and “Marital Status – Married” from our attributes to deliver recommendations on. Statistically we are unable to verify there is a link to Attrition, whilst there may be a small association(Refer to Significance Codes below model outputs)

```
Best predictors: BusinessTravel + JobInvolvement + MaritalStatus + NumCompaniesWorked + OverTime + TotalW
rkingYears + YearsSinceLastPromotion
> message("Cross-validation score: ", cv_score(best_predictors))
Cross-validation score: 0.788030083899308
>
>
> final_model <- glm(
+   # Build the formula from the predictors
+   formula = paste("Attrition ~" , paste(best_predictors, collapse = " + "), sep = ""),
+   data = employees,
+   family = "binomial"
+ )
>
> summary(final_model)

Call:
glm(formula = paste("Attrition ~", paste(best_predictors, collapse = " + "),
    sep = ""), family = "binomial", data = employees)

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -1.60736    0.59485  -2.702 0.006890 **
BusinessTravelTravel_Frequently  1.27747    0.41871   3.051 0.002281 **
BusinessTravelTravel_Rarely  0.47477    0.39355   1.206 0.227674
JobInvolvement              -0.58116    0.13613  -4.269 1.96e-05 ***
MaritalStatusMarried         0.58671    0.30457   1.926 0.054057 .
MaritalStatusSingle          1.51086    0.30370   4.975 6.53e-07 ***
NumCompaniesWorked           0.15071    0.03913   3.852 0.000117 ***
OverTimeYes                  1.46253    0.19841   7.371 1.69e-13 ***
TotalWorkingYears           -0.12607    0.01996  -6.316 2.68e-10 ***
YearsSinceLastPromotion      0.13887    0.03780   3.674 0.000239 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 885.94  on 999  degrees of freedom
Residual deviance: 704.80  on 990  degrees of freedom
AIC: 724.8

Number of Fisher Scoring iterations: 6
```

Appendix D – Predictive Model Two Decision Tree Outputs

“Best Model”

Accuracy - AUC Score 0.7494
Number of Predictors – 11

Method

Decision TreeRegression using Ten-Fold Cross Validation and Backward Stepwise Selection

Backward Subset Selection was chosen with a view it has the best chance of the most important variables being in the final model, and in our trials have proven to have a better AUC score and more significant predictor variables as outputs in the models we have run.

```
Best predictors: Department, DistanceFromHome, MonthlyIncome, NumCompaniesWorked, OverTime, RelationshipSa
tisfaction, StockOptionLevel, TotalWorkingYears, TrainingTimesLastYear, YearsInCurrentRole, YearsWithCurrM
anager
> message("Cross-validation score: ", cv_score(best_predictors))
Cross-validation score: 0.74938195586061
> model
n= 1000

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 1000 162 0 (0.83800000 0.16200000)
   2) TotalWorkingYears>=2.5 919 122 0 (0.86724701 0.13275299)
     4) OverTime=No 660  57 0 (0.91363636 0.08636364) *
     5) OverTime=Yes 259  65 0 (0.74903475 0.25096525)
      10) MonthlyIncome>=3924 174  26 0 (0.85057471 0.14942529)
        20) StockOptionLevel>=0.5 109   7 0 (0.93577982 0.06422018) *
        21) StockOptionLevel< 0.5 65  19 0 (0.70769231 0.29230769)
          42) Department=Human Resources,Research & Development 40   5 0 (0.87500000 0.12500000) *
          43) Department=Sales 25  11 1 (0.44000000 0.56000000)
            86) MonthlyIncome< 8163.5 15   6 0 (0.60000000 0.40000000) *
            87) MonthlyIncome>=8163.5 10   2 1 (0.20000000 0.80000000) *
      11) MonthlyIncome< 3924 85  39 0 (0.54117647 0.45882353)
        22) TrainingTimesLastYear< 3.5 66  26 0 (0.60606061 0.39393939)
          44) MonthlyIncome>=2457 41  12 0 (0.70731707 0.29268293) *
          45) MonthlyIncome< 2457 25  11 1 (0.44000000 0.56000000)
            90) TotalWorkingYears>=8.5 7   2 0 (0.71428571 0.28571429) *
            91) TotalWorkingYears< 8.5 18   6 1 (0.33333333 0.66666667) *
        23) TrainingTimesLastYear>=3.5 19   6 1 (0.31578947 0.68421053) *
   3) TotalWorkingYears< 2.5 81  40 0 (0.50617284 0.49382716)
     6) OverTime=No 53  20 0 (0.62264151 0.37735849)
      12) Department=Research & Development 35   8 0 (0.77142857 0.22857143) *
      13) Department=Human Resources,Sales 18   6 1 (0.33333333 0.66666667) *
     7) OverTime=Yes 28   8 1 (0.28571429 0.71428571) *
>
```

"Chosen Model"

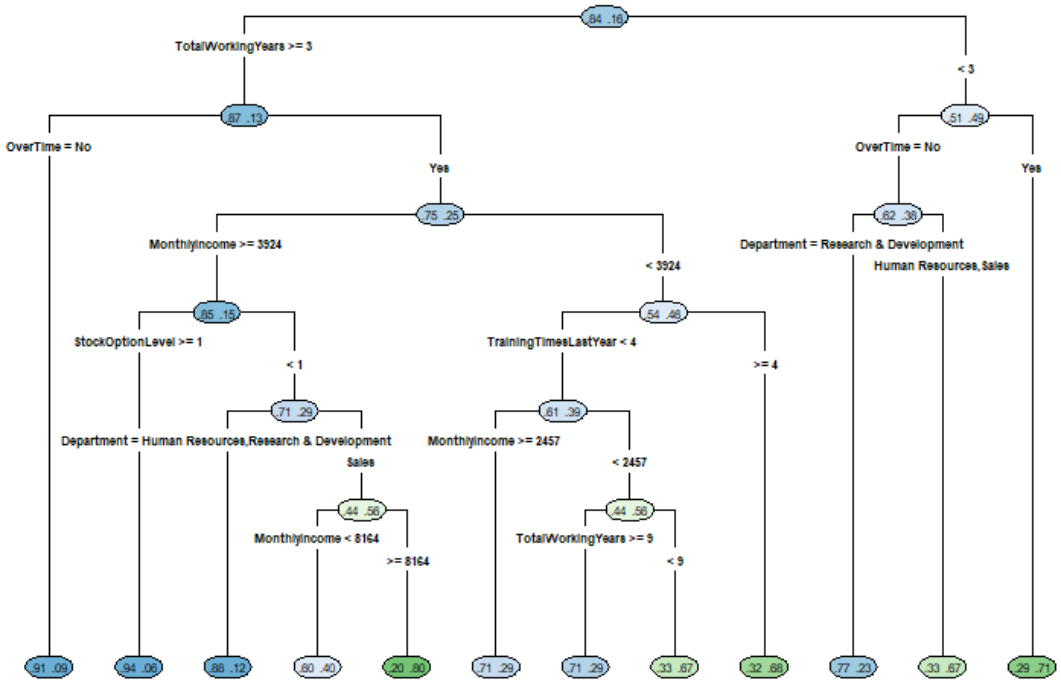
Accuracy - AUC Score 0.7279 accepting a 0.01 reduction in AUC for simpler model
Number of Predictors - 6
Method – Decision Tree using Ten-Fold Cross Validation Backward Stepwise Selection

```
> message("Best predictors: ", paste(best_predictors, collapse = ", "))
Best predictors: Department, MonthlyIncome, OverTime, StockOptionLevel, TotalWorkingYears, TrainingTimesLa
stYear
> message("Cross-validation score: ", cv_score(best_predictors))
Cross-validation score: 0.72791939281056
```

```
n= 1000

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 1000 162 0 (0.83800000 0.16200000)
   2) TotalWorkingYears>=2.5 919 122 0 (0.86724701 0.13275299)
     4) OverTime=No 660  57 0 (0.91363636 0.08636364) *
     5) OverTime=Yes 259  65 0 (0.74903475 0.25096525)
      10) MonthlyIncome>=3924 174  26 0 (0.85057471 0.14942529)
        20) StockOptionLevel>=0.5 109   7 0 (0.93577982 0.06422018) *
        21) StockOptionLevel< 0.5 65  19 0 (0.70769231 0.29230769)
          42) Department=Human Resources,Research & Development 40   5 0 (0.87500000 0.12500000) *
          43) Department=Sales 25  11 1 (0.44000000 0.56000000)
            86) MonthlyIncome< 8163.5 15   6 0 (0.60000000 0.40000000) *
            87) MonthlyIncome>=8163.5 10   2 1 (0.20000000 0.80000000) *
      11) MonthlyIncome< 3924 85  39 0 (0.54117647 0.45882353)
        22) TrainingTimesLastYear< 3.5 66  26 0 (0.60606061 0.39393939)
          44) MonthlyIncome>=2457 41  12 0 (0.70731707 0.29268293) *
          45) MonthlyIncome< 2457 25  11 1 (0.44000000 0.56000000)
            90) TotalWorkingYears>=8.5 7   2 0 (0.71428571 0.28571429) *
            91) TotalWorkingYears< 8.5 18   6 1 (0.33333333 0.66666667) *
        23) TrainingTimesLastYear>=3.5 19   6 1 (0.31578947 0.68421053) *
   3) TotalWorkingYears< 2.5 81  40 0 (0.50617284 0.49382716)
     6) OverTime=No 53  20 0 (0.62264151 0.37735849)
      12) Department=Research & Development 35   8 0 (0.77142857 0.22857143) *
      13) Department=Human Resources,Sales 18   6 1 (0.33333333 0.66666667) *
     7) OverTime=Yes 28   8 1 (0.28571429 0.71428571) *
> best_predictors
[1] "Department"            "MonthlyIncome"         "OverTime"              "StockOptionLevel"
[5] "TotalWorkingYears"     "TrainingTimesLastYear"
```



Appendix E - R Code

Predictive Model One - Logistic Regression

```
# Load the libraries we need
suppressMessages(library(pROC))


#Import employees2.csv
employees <- read.csv('employees2.csv')

# Prepare the data
employees$Attrition = ifelse(employees$Attrition == "Yes", 1, 0)

# Create the ten folds
set.seed(57)
employees$Fold <- sample(rep(1:10, length.out=nrow(employees)))

# Define a function to do the cross-validation
cv_score <- function(predictors) {
  scores <- c()
  for (fold in 1:10) {
    training = employees[employees$Fold != fold,]
    validation = employees[employees$Fold == fold,]
    model <- glm(
      # Build the formula from the predictors
      formula = paste("Attrition ~ ", paste(predictors, collapse = "
+ "), sep = ""),
      data = training,
      family = "binomial"
```

```
    )
    validation$Probability <- predict(
      object = model,
      newdata = validation,
      type = "response"
    )
    auc <- roc(validation$Attrition ~ validation$Probability,
quiet=TRUE)$auc
    scores <- c(scores, auc)
  }
  return(mean(scores))
}

#---------------LOOK FOR BEST AUC SCORE POSSIBLE --------


# Define a function to do the dropping of predictor variables
step_backward_best <- function(current_predictors) {

  # If there aren't at least two current predictors, return the
current_predictors
  if (length(current_predictors) <= 1) return(current_predictors)

  # Get the cross-validation score of the current predictors
  current_score <- cv_score(current_predictors)

  # Show the current predictors and their cross-validation score
  message("Current predictors: ", paste(current_predictors,
collapse=" + "))
  message("Current score: ", current_score)

  # Create a vector in which to keep the new scores
  new_scores <- c()

  # Try dropping each of the current predictors
  for (predictor_to_try in current_predictors) {
    # Remove the predictor from the current predictors
    new_predictors <- current_predictors[current_predictors !=
predictor_to_try]
    # Find the cross-validation score of the new predictors
    new_score <- cv_score(new_predictors)
    # Show the score
    message("Without ", predictor_to_try, ": ", new_score)
    # Add it to the vector of new scores
    new_scores[predictor_to_try] <- new_score
  }

  # If the best new score is at least as good as the current score
  if (max(new_scores) >= current_score) {
    # Find the predictor to drop
    predictor_to_drop <- names(new_scores)[which.max(new_scores)]
    # Output an explanatory message
    message("Dropping ", predictor_to_drop, "\n")
    # Remove it from the current predictors
    current_predictors <- current_predictors[current_predictors !=
predictor_to_drop]
```

```
    # Step backward again
    return(step_backward_best(current_predictors))
  }

  # Otherwise
  else {
    message("Nothing to drop\n")
    return(current_predictors)
  }
}

# Get the starting predictors
starting_predictors <- colnames(employees)
starting_predictors <- starting_predictors[starting_predictors !=
"Attrition" & starting_predictors != "Fold"]

# Apply the stepping process looking for highest AUC score
best_predictors <- step_backward_best(starting_predictors)

# Show the final results
message("Best predictors highest AUC: ", paste(best_predictors,
collapse = ", "))
message("Cross-validation score: ", cv_score(best_predictors))

#---------------LOOK FOR LOWER AUC SCORE WITH SIMPLER MODEL --------

# Define a function to do the dropping of predictor variables
step_backward_simple <- function(current_predictors) {

  # If there aren't at least two current predictors, return the
current_predictors
  if (length(current_predictors) <= 1) return(current_predictors)

  # Get the cross-validation score of the current predictors
  current_score <- cv_score(current_predictors)

  # Show the current predictors and their cross-validation score
  message("Current predictors: ", paste(current_predictors,
collapse=" + "))
  message("Current score: ", current_score)

  # Create a vector in which to keep the new scores
  new_scores <- c()

  # Try dropping each of the current predictors
  for (predictor_to_try in current_predictors) {
    # Remove the predictor from the current predictors
    new_predictors <- current_predictors[current_predictors !=
predictor_to_try]
    # Find the cross-validation score of the new predictors
    new_score <- cv_score(new_predictors)
    # Show the score
    message("Without ", predictor_to_try, ": ", new_score)
    # Add it to the vector of new scores
    new_scores[predictor_to_try] <- new_score
  }
```

```
  # If the best new score is at least 0.007 less than the current
score
  if (max(new_scores) >= current_score - 0.007) {
    # Find the predictor to drop
    predictor_to_drop <- names(new_scores)[which.max(new_scores)]
    # Output an explanatory message
    message("Dropping ", predictor_to_drop, "\n")
    # Remove it from the current predictors
    current_predictors <- current_predictors[current_predictors !=
predictor_to_drop]
    # Step backward again
    return(step_backward_simple(current_predictors))
  }

  # Otherwise
  else {
    message("Nothing to drop\n")
    return(current_predictors)
  }
}

# Define the starting predictors
starting_predictors <- colnames(employees[!colnames(employees) %in%
c("Attrition", "Fold")])


# Use step_backward to find the best predictors
best_predictors_simple <- step_backward_simple(starting_predictors)

# Show the results
message("Best predictors, simple model: ",
paste(best_predictors_simple, collapse = " + "))
message("Cross-validation score: ",
cv_score(best_predictors_simple))


#Build the final model with teh simpler version of predictors
final_model <- glm(
  # Build the formula from the predictors
  formula = paste("Attrition ~" , paste(best_predictors_simple,
collapse = " + "), sep = ""),
  data = employees,
  family = "binomial"
)

#Output Model
summary(final_model)
```

Predictive Model Two - Decision Tree

```
# Import Libraries
suppressMessages(library(rpart))
suppressMessages(library(rpart.plot))
```

```
suppressMessages(library(pROC))

# Load and prepare the data
employees <- read.csv(file="employees2.csv")

# Set up for Cross Validation, creating 10 folds in the dataset
employees$Attrition <- ifelse(employees$Attrition == "Yes", 1, 0)
set.seed(57)
employees$Fold <- sample(rep(1:10, length.out=nrow(employees)))

# Define a function to do the cross-validation
cv_score <- function(predictors) {
  scores <- c()
  # For each fold, train the model on the data on the other nine
folds,
  # Validate on the current fold
  for (fold in 1:10) {
    training = employees[employees$Fold != fold,]
    validation = employees[employees$Fold == fold,]
    model <- rpart(
      formula = paste("Attrition ~ ", paste(predictors, collapse = "
+ "), sep = ""),
      data = training,
      method = "class"
    )
    validation$Probability <- predict(
      object = model,
      newdata = validation,
      type = "prob"
    )
    # Generate the AUC score(accuracy) of the model using the
current fold
    auc <- roc(validation$Attrition ~ validation$Probability[ , 2],
quiet=TRUE)$auc
    scores <- c(scores, auc)
  }
  # Return the mean accuracy/AUC score for all the models
  return(mean(scores))
}


#---------------LOOK FOR BEST AUC SCORE POSSIBLE --------


# Define a function to do the dropping of predictor variables
step_backward_best <- function(current_predictors) {
  if (length(current_predictors) <= 1) return(current_predictors)
  current_score <- cv_score(current_predictors)
  message("Current predictors: ", paste(current_predictors,
collapse=", "))
  message("Current score: ", current_score)
  new_scores <- c()
  for (predictor in current_predictors) {
    new_predictors <- current_predictors[current_predictors !=
predictor]
    new_score <- cv_score(new_predictors)
```

```
    message("Without ", predictor, ": ", new_score)
    new_scores[predictor] <- new_score
  }
  if (max(new_scores) >= current_score ) {
    predictor_to_drop <- names(new_scores)[which.max(new_scores)]
    message("Dropping ", predictor_to_drop, "\n")
    new_predictors <- current_predictors[current_predictors !=
predictor_to_drop]
    return(step_backward_best(new_predictors))
  }
  else {
    message("Nothing to drop\n")
    return(current_predictors)
  }
}

# Get the starting predictors
starting_predictors <- colnames(employees)
starting_predictors <- starting_predictors[starting_predictors !=
"Attrition" & starting_predictors != "Fold"]

# Apply the stepping process looking for highest AUC score
best_predictors <- step_backward_best(starting_predictors)

# Show the final results
message("Best predictors highest AUC: ", paste(best_predictors,
collapse = ", "))
message("Cross-validation score: ", cv_score(best_predictors))

#--------LOOK FOR LOWER AUC SCORE BY 0.01 FOR A SIMPLER MODEL ------


# Modify above function to do the dropping
step_backward_simple <- function(current_predictors) {
  if (length(current_predictors) <= 1) return(current_predictors)
  current_score <- cv_score(current_predictors)
  message("Current predictors: ", paste(current_predictors,
collapse=", "))
  message("Current score: ", current_score)
  new_scores <- c()
  for (predictor in current_predictors) {
    new_predictors <- current_predictors[current_predictors !=
predictor]
    new_score <- cv_score(new_predictors)
    message("Without ", predictor, ": ", new_score)
    new_scores[predictor] <- new_score
  }
  # Accept an AUC score 0.01 lower if a variable can be dropped
  if (max(new_scores) >= current_score - 0.01) {
    predictor_to_drop <- names(new_scores)[which.max(new_scores)]
    message("Dropping ", predictor_to_drop, "\n")
    new_predictors <- current_predictors[current_predictors !=
predictor_to_drop]
    return(step_backward_simple(new_predictors))
  }
  else {
```

```
    message("Nothing to drop\n")
    return(current_predictors)
  }
}

# Get the starting predictors
starting_predictors <- colnames(employees)
starting_predictors <- starting_predictors[starting_predictors !=
"Attrition" & starting_predictors != "Fold"]

# Apply the stepping process looking for a simpler model
best_predictors_simple <- step_backward_simple(starting_predictors)

# Show the final results
message("Best predictors, simpler model: ",
paste(best_predictors_simple, collapse = ", "))
message("Cross-validation score: ",
cv_score(best_predictors_simple))

# Train the model on the full dataset
model <- rpart(
  formula = paste("Attrition ~ ", paste(best_predictors_simple,
collapse = " + "), sep = ""),
  data = employees,
  method = "class"
)

# Print Decision Tree
rpart.plot(
  model,
  type = 5,
  extra = 7
)

# Print the Model Output
model
```

R –Code for Visualisations in Report

```
# Import Libraries
library(writexl)
suppressMessages(library(pROC))

#Import data files
employees <- read.csv('employees2.csv', stringsAsFactors = T)
dictionary <- read.csv('Dictionary.csv', header = FALSE, col.names =
c('SurveyQuestion', 'QuestionDetail'))

# Write Dictionary to an excel file for report
write_xlsx(dictionary,'Dictionary.xlsx')

# Set parameters for plots
par(
```

```
  # Specify the margins: bottom, left, top, right
  mar = c(6, 10, 10, 6),
  cex.axis = 0.9
)


# Create a barplot of employees that have left and stayed
mids <- barplot(table(employees$Attrition),
        main = "Number of Employees Who Have Left Globex \n Since
2023 Employee Survey",
        ylim = c(0,1000),
        col = c("lightblue", "indianred1"),
        las = 1,
        names.arg = c("Stayed", "Left")
)

# Place count data of number employees who haveleft and stayed on
the bars
text(
  x = mids,
  y = table(employees$Attrition),
  labels = table(employees$Attrition),
  pos = 3)

# Specify the margins: bottom, left, top, right
par(
  mar = c(6, 11, 4.1, 2.1)
)


# Create a barplot of Overtime in the organisaton
mids2 <- barplot(
    table(employees$OverTime),
    main = "Overtime at Globex",
    ylim = c(0,1000),
    col = c("lightblue", "indianred1"),
    las = 1,
    names.arg = c("No", "Yes")
)

# create labels of number of employees on bars
text(
  x = mids2,
  y = table(employees$OverTime),
  labels = table(employees$OverTime),
  pos = 3)

write_xlsx(income_by_job[order(income_by_job$MonthlyIncome),],'Month
lyMedianIncome.xlsx')

#Create data of employees Working Years less than 3 years
junior_employees <- aggregate(
  Department ~ TotalWorkingYears,
  data = employees[employees$TotalWorkingYears < 3,],
  FUN = length)
junior_employees
```

```
# Create Plot of number of mployees in Yerars 1 - 3
mids6 <- barplot(
  junior_employees$Department ~ junior_employees$TotalWorkingYears,
  main = "WorkingYears <3 @ Globex \n",
  col = c("lightblue2", "dodgerblue","cadetblue"),
  names.arg = c("First Year", "Second Year", "Third Year"),
  ylab = "Number of Employees",
  xlab = "",
  las = 1
  )

# Create labels
text(
  x = mids6,
  y = junior_employees$Department,
  labels = junior_employees$Department,
  pos = 3)

# Create data for Attrition By Role Employees less than 3 years
junior_employees2 <- prop.table(
  table(junior_employees_info$Attrition,
junior_employees_info$JobRole),
  margin = 2)

junior_employees2 <- junior_employees2[,colnames(junior_employees2)
%in% c("Human Resources", "Sales Representative", "Laboratory
Technician", "Research Scientist")]
junior_employees2 <-
junior_employees2[,order(junior_employees2[2,])]

par(
  mar = c(6, 2, 10, 6)
)

# create a barplot of the Working Years Table
mids3 <- barplot(
  junior_employees2,
  main = "Total Working Years Less Than 3 Years \n Attrition By
Role",
  xlab = "",
  yaxt= "n",
  col = c("lightblue2", "indianred1"),
  las = 1,
  beside = T,
  ylim = c(0,.8),
  cex.lab = 1
)

# Create labels for plot
text(
  x = mids3,
  y = junior_employees2,
  labels = paste(round(junior_employees2,2) * 100, "%"),
  pos = 3)
```

```
# Create a legend
legend(
  # Put at the top
  "topright",
  xpd = T,
  inset = c(-.1, 0),
  # Specify the labels in the legend
  legend = c("Stayed", "Left"),
  # Specify the colours in the legend
  fill = c("cadetblue1","indianred1"),
  # Remove box from legend
  bty = "n"
)

median(employees$MonthlyIncome)

junior_wages <- subset(employees, employees$TotalWorkingYears < 3)

median(junior_wages$MonthlyIncome)

# Create data for Attrition By Business Treval
travel_vs_leavers <- prop.table(table(employees$Attrition,
employees$BusinessTravel), margin = 2)
travel_vs_leavers <-
travel_vs_leavers[,order(travel_vs_leavers[2,])]

par(
  mar = c(4,4,4,4)
)

# Create Barplot
mids4 <- barplot(
  travel_vs_leavers[,order(travel_vs_leavers[2,])],
  las = 1,
  beside = T,
  yaxt = "n",
  width = c(10,9),
  col = c("lightblue", "indianred1"),
  ylim = c(0,1),
  main = "Percentage of Employees Left By Business Travel"
)

# Create labels for plot
text(
  y = travel_vs_leavers[,order(travel_vs_leavers[2,])],
  x = mids4,
  labels = paste(round(travel_vs_leavers,2) * 100, "%"),
  pos = 3
  )

# Create a legend
legend(
  # Put at the top
  "topright",
  # Specify the labels in the legend
  legend = c("Stayed", "Left"),
```

```
  # Specify the colours in the legend
  fill = c("cadetblue1","indianred1"),
  # Remove box from legend
  bty = "n"
)


par(
  mar = c(8,4,4,4)
)

# Get data for employees over $8164 Monthly Income
highincome_leavers <- subset(employees, employees$MonthlyIncome >
8164)
nostock_vs_leavers <- prop.table(
  table(highincome_leavers$Attrition,
highincome_leavers$StockOptionLevel),
        margin = 2)

# Create a plot Attrition By Stock Option leve, Monthly Income $8164
mids5 <- barplot(
  nostock_vs_leavers,
  las = 1,
  beside = T,
  yaxt = "n",
  width = c(10,9),
  col = c("lightblue", "indianred1"),
  ylim = c(0,1),
  names.arg = c("None", "Some", "High", "Very High"),
  main = "Percentage of Leavers \n Income > $8164 vs. Stock Option
Level",
)

# create labels for plot
text(
  y = nostock_vs_leavers,
  x = mids5,
  labels = paste(round(nostock_vs_leavers,2) * 100, "%"),
  pos = 3
)

# Create a legend
legend(
  # Put at the top
  "topright",
  xpd = T,
  inset = c(-.1, 0),
  # Specify the labels in the legend
  legend = c("Stayed", "Left"),
  # Specify the colours in the legend
  fill = c("cadetblue1","indianred1"),
  # Remove box from legend
  bty = "n"
)
```