Attrition Project

Background:

The workplace is currently going through, what most call, the Great Resignation. Jobs are becoming more competitive than ever before and everyone wants to have the highest talent they can find, but with this many employees that have shown their skills and proved their value tend to have options and the ability to easily take their talents elsewhere. Due to this, many companies have seen attrition rise to record levels and are doing everything they can to combat it.

Objective:

We will be doing a deep dive into attrition and what are the leading factors that cause people to leave; is it job satisfaction? Is it just salary differences? Or are there other variables that come into the equation for different sets of people? We are going to figure out how best this company can improve to best satisfy their employees and try to retain and attract the best talent. We are also aiming to figure out what the thresholds are for the most important variables before an employee is going to leave. This will help our company be able to prioritize the needs of employees and satisfy them before they decide to leave the company.

Data Set:

We will be using a dataset from an anonymous company that has 2,940 employee's data that covers 34 different variables from their daily pay rate, age and education all the way to their amount of their commutes and how they rate their work life balance.

The most important is that this dataset has the response variable of attrition, which is marked yes if the employee has left the company and no if they are still currently working there. We will denote 1 if the employee has left and 0 if the employee has not left the company.

Variables:

1. EmployeeNumber - Employee Identifier
2. Attrition - Did the employee attrite? (response variable)
3. Age - Age of the employee
4. BusinessTravel - Travel commitments for the job (categorical)
5. DailyRate - Data description not available**
6. Department - Employee Department (categorical)
7. DistanceFromHome - Distance from work to home (in km)
8. Education - 1-Below College, 2-College, 3-Bachelor, 4-Master,5-Doctor (categorical)
9. EducationField - Field of Education (categorical)
10. EnvironmentSatisfaction - 1-Low, 2-Medium, 3-High, 4-Very High (categorical)
11. Gender - Employee's gender (categorical)
12. HourlyRate - Data description not available**
13. JobInvolvement - 1-Low, 2-Medium, 3-High, 4-Very High (categorical)
14. JobLevel - Level of job (1 to 5)
15. JobRole - Job Roles (categorical)
16. JobSatisfaction - 1-Low, 2-Medium, 3-High, 4-Very High (categorical)
17. MaritalStatus - Marital Status (categorical)

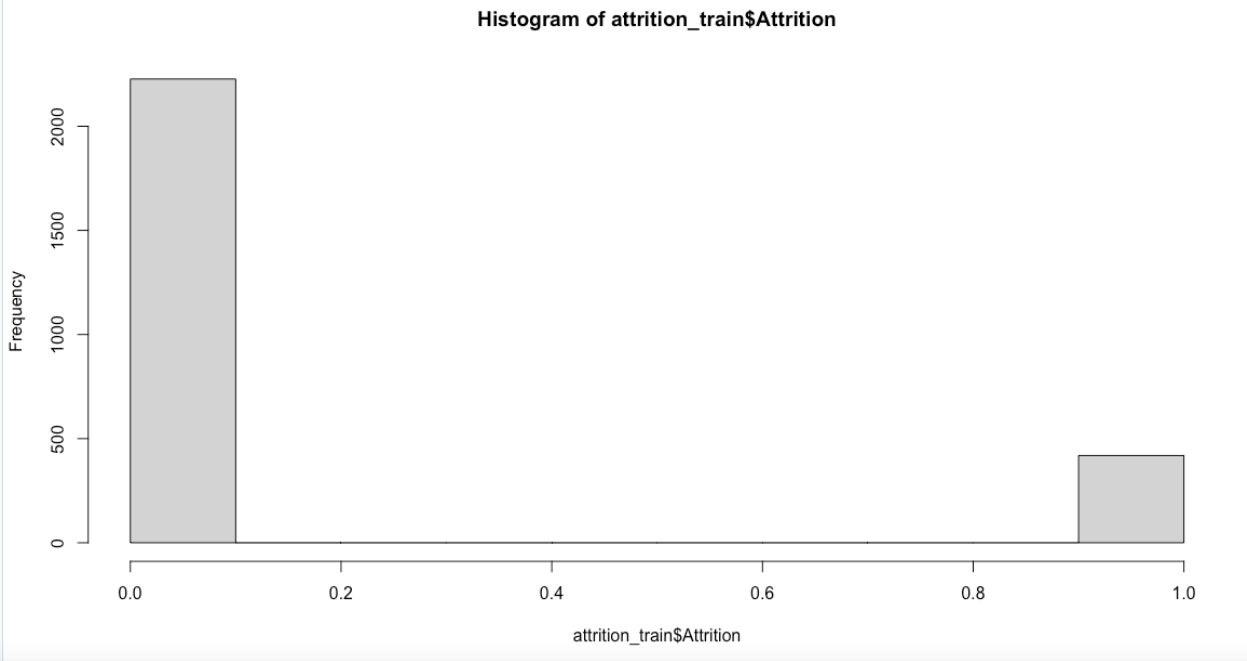
18. MonthlyIncome - Monthly Salary
19. MonthlyRate - Data description not available**
20. NumCompaniesWorked - Number of companies worked at
21. Over18 - Over 18 years of age? (categorical)
22. OverTime - Overtime? (categorical)
23. PercentSalaryHike - The percentage increase in salary last year
24. PerformanceRating - 1-Low, 2-Good, 3-Excellent, 4-Outstanding (categorical)
25. RelationshipSatisfaction - 1-Low, 2-Medium, 3-High, 4-Very High (categorical)
26. StandardHours - Standard Hours
27. StockOptionLevel - Stock Option Level
28. TotalWorkingYears - Total years worked
29. TrainingTimesLastYear - Number of training attended last year
30. WorkLifeBalance - 1-Low, 2-Good, 3-Excellent, 4-Outstanding
31. YearsAtCompany - Years at Company
32. YearsInCurrentRole - Years in the current role
33. YearsSinceLastPromotion - Years since the last promotion
34. YearsWithCurrManager - Years with the current manager

Exploratory data analysis:

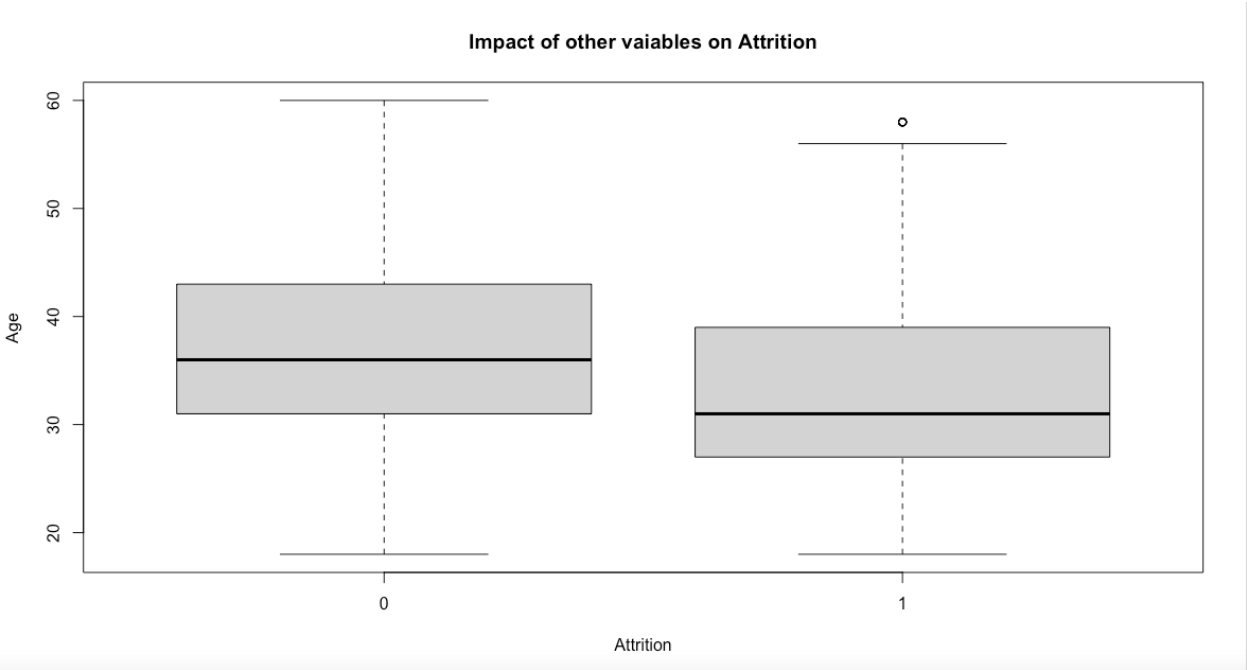
```
Call:
glm(formula = Attrition ~ ., family = binomial, data = attrition_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6826  -0.4753  -0.2435  -0.0814   3.4819

Coefficients: (2 not defined because of singularities)
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -1.063e+01  4.749e+02  -0.022 0.982138
EmployeeNumber                  -7.983e-05  7.897e-05  -1.011 0.312070
Age                             -3.783e-02  1.041e-02  -3.635 0.000278 ***
BusinessTravelTravel_Frequently  1.916e+00  3.056e-01   6.268 3.65e-10 ***
BusinessTravelTravel_Rarely      9.654e-01  2.788e-01   3.462 0.000536 ***
DailyRate                       -3.109e-04  1.666e-04  -1.867 0.061942 .
DepartmentResearch & Development 1.386e+01  4.749e+02   0.029 0.976714
DepartmentSales                  1.385e+01  4.749e+02   0.029 0.976735
DistanceFromHome                 4.841e-02  8.307e-03   5.828 5.61e-09 ***
Education                        2.490e-02  6.695e-02   0.372 0.709906
EducationFieldLife Sciences     -1.099e+00  6.041e-01  -1.819 0.068887 .
EducationFieldMarketing         -5.822e-01  6.421e-01  -0.907 0.364517
EducationFieldMedical           -1.197e+00  6.047e-01  -1.979 0.047774 *
EducationFieldOther             -9.235e-01  6.423e-01  -1.438 0.150492
EducationFieldTechnical Degree  -1.071e-01  6.157e-01  -0.174 0.861915
EnvironmentSatisfaction         -4.850e-01  6.355e-02  -7.631 2.34e-14 ***
GenderMale                       4.306e-01  1.406e-01   3.064 0.002184 **
HourlyRate                       1.648e-03  3.352e-03   0.492 0.623007
JobInvolvement                  -5.118e-01  9.265e-02  -5.524 3.31e-08 ***
JobLevel                        -1.615e-01  2.437e-01  -0.663 0.507417
JobRoleHuman Resources           1.502e+01  4.749e+02   0.032 0.974771
JobRoleLaboratory Technician     1.448e+00  3.637e-01   3.981 6.85e-05 ***
JobRoleManager                   1.146e-01  6.786e-01   0.169 0.865952
JobRoleManufacturing Director    3.439e-01  3.980e-01   0.864 0.387505
JobRoleResearch Director        -1.132e+00  7.261e-01  -1.559 0.118906
JobRoleResearch Scientist        5.252e-01  3.707e-01   1.417 0.156571
JobRoleSales Executive           9.739e-01  8.353e-01   1.166 0.243654
JobRoleSales Representative      1.859e+00  8.797e-01   2.113 0.034560 *
```

```
Over18                          NA         NA      NA       NA
OverTime                  2.014e+00  1.481e-01  13.599  < 2e-16 ***
PercentSalaryHike        -2.132e-02  2.988e-02  -0.713 0.475542
PerformanceRating         1.455e-01  3.027e-01   0.481 0.630695
RelationshipSatisfaction -3.166e-01  6.279e-02  -5.043 4.58e-07 ***
StandardHours                   NA         NA      NA       NA
StockOptionLevel         -1.932e-01  1.217e-01  -1.587 0.112488
TotalWorkingYears        -7.076e-02  2.252e-02  -3.142 0.001679 **
TrainingTimesLastYear    -2.058e-01  5.524e-02  -3.726 0.000194 ***
WorkLifeBalance          -4.155e-01  9.369e-02  -4.435 9.20e-06 ***
YearsAtCompany            9.985e-02  3.029e-02   3.297 0.000978 ***
YearsInCurrentRole       -1.740e-01  3.533e-02  -4.925 8.44e-07 ***
YearsSinceLastPromotion   1.844e-01  3.320e-02   5.555 2.78e-08 ***
YearsWithCurrManager     -1.333e-01  3.623e-02  -3.679 0.000234 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2312.2  on 2645  degrees of freedom
Residual deviance: 1496.4  on 2600  degrees of freedom
AIC: 1588.4

Number of Fisher Scoring iterations: 15
```

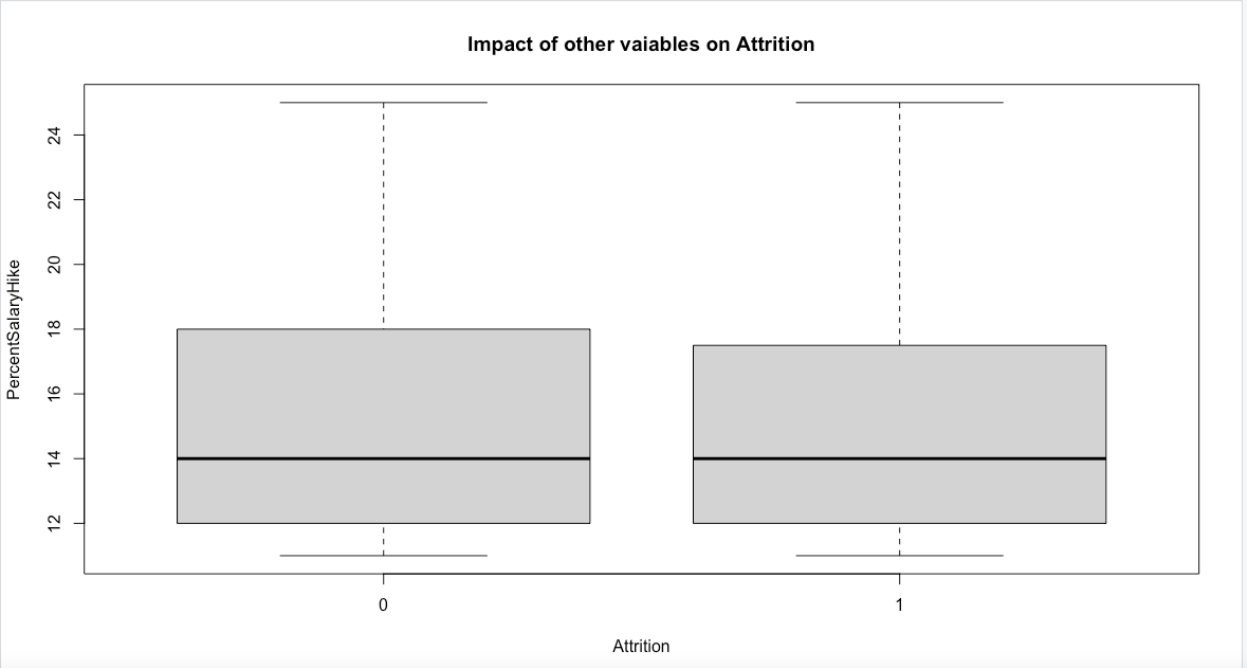
After we evaluate the goodness of fit, we can see that most of the p-values are relatively small compared to the confidence level. The summary statistics also tells us that the coefficients of variables Over18 and StandardHours are not defined because they have a perfect linear relationship. Also, since the response variable that we chose is a binary variable so the model we used in this case is a logistic regression. We also noted the star levels on the right hand side, marking which variables this model believes to be strong indicators.



We also looked at the histogram of response variable Attrition and found that this company's attrition rate is sitting around 16.12% and the average "good" attrition rate that companies try to aim for is below 10%. This shows that there is some area to improve and we need to dig further to see why these employees are leaving .



This box plot tells us that most employees that have not left tend to be older, between the ages of 30 and 45. The people who have chosen to take their talents elsewhere are on average younger, mostly younger than 40. While there are definitely outliers on both sides the boxes do show that younger people tend to be more willing to leave.



We also found another pattern after looking at the second box plot which shows us that most employees that have not left their company usually have slightly higher salary raises compared to those who left. We expected that there would be more of a difference in these boxes, pointing to the fact that pay/raises would be a key indicator. One thing to note is that this is a percentage of their raise in the previous year and does not take into account the actual pay numbers, but just a percentage. This boxplot makes us think that these people are not just leaving due to issues with pay, but other reasons.

Logistic Regression

Stepwise type	**AIC**
Backwards Stepwise	1638.056
Forwards Stepwise	1533.865

For both forwards and backwards we attained the same 22 most important variables, but because we found that the AIC for the forwards stepwise model was lower we decided that it would be our best model going forward.

Below are the 22 most important variables that were found by the forward stepwise logistic regression. They are in order of most important to the model to least important. From this we can see that over time and the employees role are the most important factors that affect attrition. Later we will dive deeper into how each one affects attrition when they are lower and higher.

1-Over TIme	2-Job Role	3-Marital Status	4-Environment Satisfaction	5- Job Involvement	6-Job Satisfaction
7-Business Travel	8-Years With Current Manager	9-Education Field	10-Distance From Home	11-Age	12-Number of Companies Worked
13-Years Since Last Promotion	14-Work Life Balance	15-Relationship Satisfaction	16-Gender	17-Training Times Last Year	18-Years In Current Role
19-Years At Company	20-Total Working years	21-Stock Option Level	22-Daily Rate		

Bootstrap

```
Bootstrap Statistics :
         original        bias     std. error
t1*   1.949231323 -7.085474e-02 0.9155154877
t2*   1.898206380  4.916196e-02 0.1483949198
t3*   1.418895416  5.369899e-02 0.4858917339
t4*   1.566673364  7.886879e-02 0.3307072741
t5*   0.109367590 -8.013514e-02 0.5797527672
t6*   0.345829054  4.413580e-02 0.4576960174
t7*  -0.892502134 -4.328174e-01 2.4841149794
t8*   0.562766605  5.347621e-02 0.3408210290
t9*   0.960944011  7.246309e-02 0.3603563047
t10*  2.106615243  1.019920e-01 0.3919218488
t11*  0.342886835  1.055023e-02 0.1884439194
t12*  1.096381475  3.864978e-02 0.2584016712
t13* -0.381788025 -9.786548e-03 0.0619734417
t14* -0.404731580 -5.619094e-03 0.0595970363
t15* -0.520339094 -1.215878e-02 0.0962101137
t16*  1.810619420  8.070194e-02 0.2863637237
t17*  0.934916989  4.983081e-02 0.2590669216
t18* -0.130450490 -3.542975e-03 0.0408348588
t19* -0.606653701 -1.151602e-02 0.4954045019
t20* -0.212410929 -6.566895e-03 0.5316413833
t21* -0.661728192 -9.919914e-03 0.5061249402
t22* -0.799192338 -2.509306e-02 0.5598002843
t23*  0.364380497  2.608853e-02 0.5281686414
t24*  0.045863274  8.382980e-04 0.0074046984
t25* -0.024491618 -3.353577e-04 0.0116029907
t26*  0.175607904  4.375796e-03 0.0274342630
t27*  0.173617061  5.832874e-03 0.0371094882
t28* -0.359386862 -6.347912e-03 0.1000088606
t29* -0.241212049 -5.442169e-03 0.0637456786
t30* -0.189006218 -2.389947e-03 0.0563982095
t31*  0.371036473  1.614273e-02 0.1403234272
t32* -0.061388449 -1.590586e-03 0.0224272940
t33* -0.137659207 -7.019940e-03 0.0339813557
t34*  0.084626276  4.016621e-03 0.0390614120
t35* -0.000354327 -1.016482e-05 0.0001682495
t36* -0.264667392 -4.776080e-04 0.1346740402
```

:

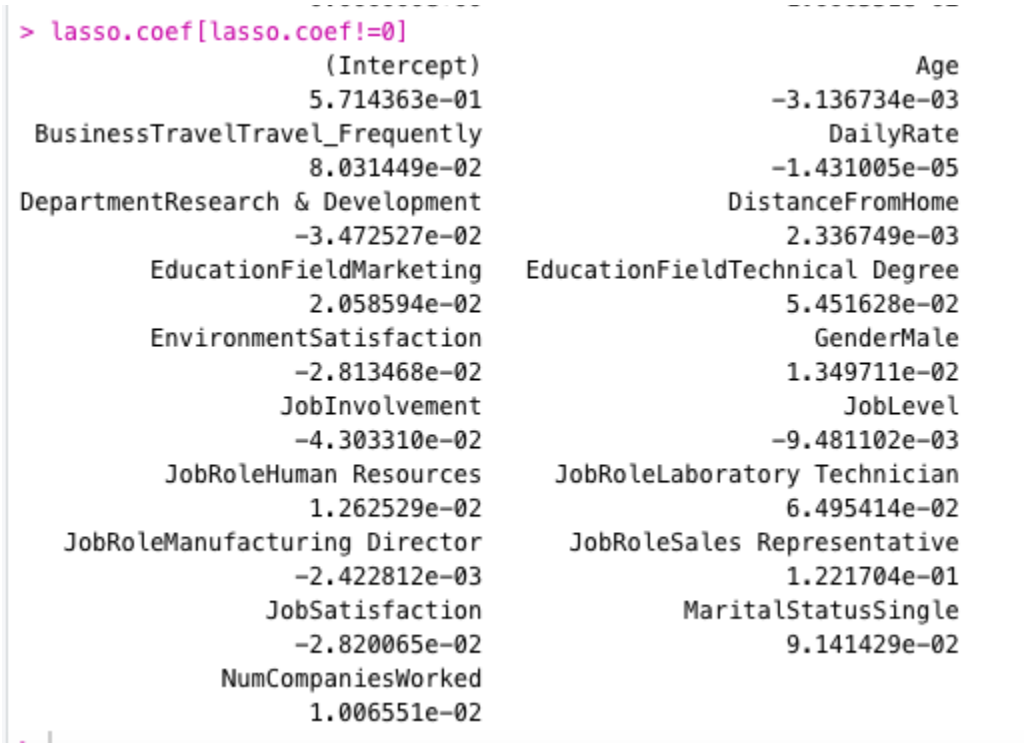
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.0476752477	0.8839321573	2.3165525	2.052812e-02
OverTime	1.9486303278	0.1450718270	13.4321761	3.917287e-41
JobRoleHuman Resources	1.4655262446	0.4848829543	3.0224330	2.507516e-03
JobRoleLaboratory Technician	1.5222947559	0.3349323553	4.5450812	5.491408e-06
JobRoleManager	0.1639904020	0.4908830847	0.3340722	7.383251e-01
JobRoleManufacturing Director	0.1540849589	0.4049687160	0.3804861	7.035846e-01
JobRoleResearch Director	-0.9008748340	0.6632246529	-1.3583253	1.743605e-01
JobRoleResearch Scientist	0.6691197952	0.3382417420	1.9782295	4.790282e-02
JobRoleSales Executive	1.1269426556	0.3393997428	3.3203993	8.988879e-04
JobRoleSales Representative	2.1666514498	0.3876720288	5.5888774	2.285422e-08
MaritalStatusMarried	0.3517649270	0.2013560320	1.7469798	8.064084e-02
MaritalStatusSingle	1.1265044127	0.2605832506	4.3230116	1.539136e-05
EnvironmentSatisfaction	-0.4577318927	0.0623295922	-7.3437332	2.077171e-13
JobSatisfaction	-0.4331483499	0.0613064317	-7.0653003	1.602690e-12
JobInvolvement	-0.5399894626	0.0917274870	-5.8868882	3.935347e-09
BusinessTravelTravel_Frequently	2.0563393969	0.3113168334	6.6052946	3.967276e-11
BusinessTravelTravel_Rarely	1.1242169473	0.2868780059	3.9187980	8.899167e-05
YearsWithCurrManager	-0.1369886345	0.0350485935	-3.9085344	9.285772e-05
EducationFieldLife Sciences	-0.5198292150	0.5622265758	-0.9245903	3.551791e-01
EducationFieldMarketing	-0.1016653192	0.5985632015	-0.1698489	8.651289e-01
EducationFieldMedical	-0.5497240918	0.5610587794	-0.9797977	3.271860e-01
EducationFieldOther	-0.6412379501	0.6227421688	-1.0297005	3.031506e-01
EducationFieldTechnical Degree	0.4329997672	0.5789395488	0.7479188	4.545091e-01
DistanceFromHome	0.0459541494	0.0080294429	5.7232052	1.045329e-08
Age	-0.0333514783	0.0100655949	-3.3134135	9.216460e-04
NumCompaniesWorked	0.2066544201	0.0287556839	7.1865590	6.644457e-13
YearsSinceLastPromotion	0.1745020461	0.0318922679	5.4716098	4.459659e-08
WorkLifeBalance	-0.3799951291	0.0931816621	-4.0780033	4.542412e-05
RelationshipSatisfaction	-0.2489496641	0.0612279150	-4.0659504	4.783711e-05
TrainingTimesLastYear	-0.1842766514	0.0552233982	-3.3369307	8.470907e-04
GenderMale	0.4175275776	0.1378076777	3.0297846	2.447282e-03
TotalWorkingYears	-0.0595658699	0.0200863644	-2.9654879	3.022033e-03
YearsInCurrentRole	-0.1548326388	0.0340154097	-4.5518381	5.317925e-06
YearsAtCompany	0.0888337382	0.0284784190	3.1193353	1.812596e-03
DailyRate	-0.0002493791	0.0001648992	-1.5123120	1.304545e-01
StockOptionLevel	-0.2250994515	0.1170506207	-1.9230949	5.446813e-02

Through analyzing the bootstrap statistics and comparing it to our logistic regression summary statistic we found that the difference between the two were very small. The average difference between the two was less than 0.016 for all of our variables. With this we confirmed that our forward stepwise logistic regression could still be considered our best model going forward.

Lasso:

Pcut value = 0.5	Asymmetric Misclassification Cost (in-sample)	Asymmetric Misclassification Cost (out-of-sample)
Forwards Selection	0.106	0.126
Lasso	0.138	0.163
Classification Tree	0.136	0.156

Here we utilized Lasso in order to check how well our model was operating in respect to the misclassification cost. As we can see the misclassification cost was much lower for our model than it was for the lasso in both in-sample and out of sample, which proves that our model is taking into account the weight and therefore creating the best model. By using the Lasso tool we were also able to find the coefficients for the variables and see how they affected the attrition equation, which we will get into below.



```
> lasso.coef[lasso.coef!=0]
                     (Intercept)                              Age
                    5.714363e-01                    -3.136734e-03
 BusinessTravelTravel_Frequently                       DailyRate
                    8.031449e-02                    -1.431005e-05
DepartmentResearch & Development                 DistanceFromHome
                   -3.472527e-02                     2.336749e-03
         EducationFieldMarketing   EducationFieldTechnical Degree
                    2.058594e-02                     5.451628e-02
         EnvironmentSatisfaction                       GenderMale
                   -2.813468e-02                     1.349711e-02
                  JobInvolvement                          JobLevel
                   -4.303310e-02                    -9.481102e-03
          JobRoleHuman Resources     JobRoleLaboratory Technician
                    1.262529e-02                     6.495414e-02
   JobRoleManufacturing Director      JobRoleSales Representative
                   -2.422812e-03                     1.221704e-01
                JobSatisfaction              MaritalStatusSingle
                   -2.820065e-02                     9.141429e-02
             NumCompaniesWorked
                    1.006551e-02
```

Variable	Coefficients	Variable	Coefficients
Age	-3.13 e-03	Business Travel Frequency	8.03 e-02
Daily Rate	-1.4 e-05	Department R&D	-3.47 e-02
Distance From Home	2.33 e-03	Environment Satisfaction	-2.881 e-02
Job Involvement	-4.30 e-02	Job Level	-9.48 e-03
Job Satisfaction	-2.82 e-02	Number of companies worked	1.00 e-02

The table above was created by analyzing the coefficients of the variables. We can see that most of the variables match what one would normally assume and help us statistically confirm those thoughts, such as attrition goes up when employees' daily rate and satisfaction go down. There is also evidence for points that are not as obvious such as younger employees tend to have high attrition rates. Below we will see some of the education majors and job roles that had stand out coefficients.

Variable	Attrition Effect	Coefficient
Marketing Degree	Higher attrition	2.05 e-02
Technical Degree	Higher Attrition	5.45 e-02
Role- Director	Lower Attrition	-2.42 e-03
Role- Sales Rep	Higher Attrition	1.22 e-01
Role- Laboratory Technician	Higher Attrition	6.49 e-02

Above we can see that employees with either a marketing or technical degree tend to have higher attrition. For the stand out roles, we found that while directors have a much lower attrition rate, sales representatives and laboratory technicians have a high attrition

rate. When we looked a little deeper into these we found that being in the sales rep or lab technician has the highest coefficients by far.

Conclusion:

We found that the top 5 contributors to our employees leaving are: amount of overtime worked, their specific role, environment satisfaction, job involvement, and job satisfaction. Marital status was also in the top, but that is out of the companies control so we decided to leave that out. Going forward we suggest that we try to focus on these factors to better our workplace and drive attrition down. Another main concern that we found was that with the job role factor, sales representatives and laboratory technicians have a much higher coefficient, which means that they are leaving at a higher rate than other positions. We advise to take a deeper dive into why people are leaving these positions at a higher rate if everything else is consistent.