A Case Study in Talent Management

by Justin Ehly

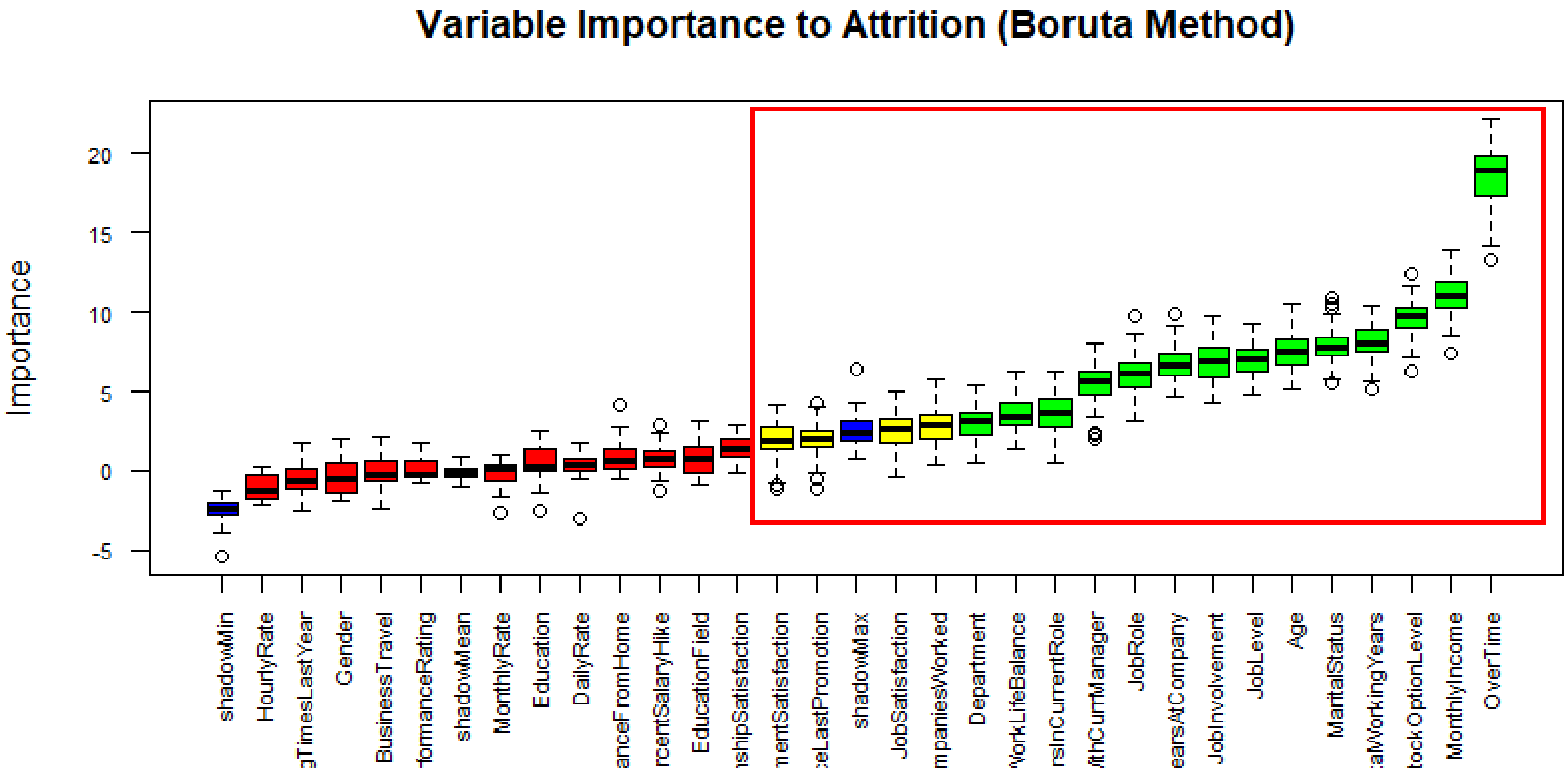
Analysis Overview

1. Attrition
2. Job Role Trends
3. Monthly Salary Estimation

Methodology

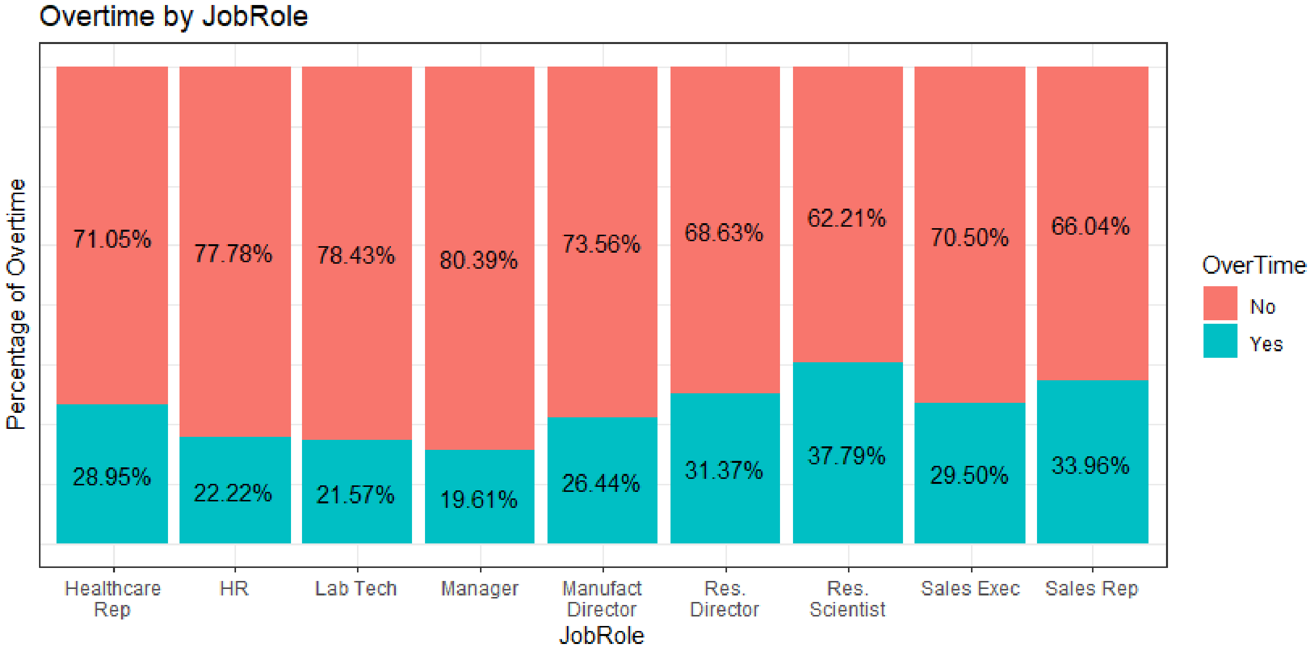
- Boruta Algorithm
  - Wrapper Built Around Random Forest Classification Algorithm
  - Top-down Search for Relevant Veatures
  - Compares Original Attributes’ Importance to Importance Achievable at Random
  - Estimates Using Permuted Copies
  - Progressively Eliminates Irrelevant Variables
  - ~ 18 Variables determined to be Significant or Possibly Significant

Attrition Variable Importance (Using Boruta)



Overtime

- 29% Employees Earn Overtime
- 71% Employees Don’t Earn Overtime
- *No Data on Bonuses
- Very low p-value vs Attrition, appears to be statistically significant
- Average Monthly Income
    - No OT: $6,464.41
    - Yes OT: $6,208.43
    - -$255.98 difference for OT Earners
- *T-tests and Chi-Sq Tests in Appendix

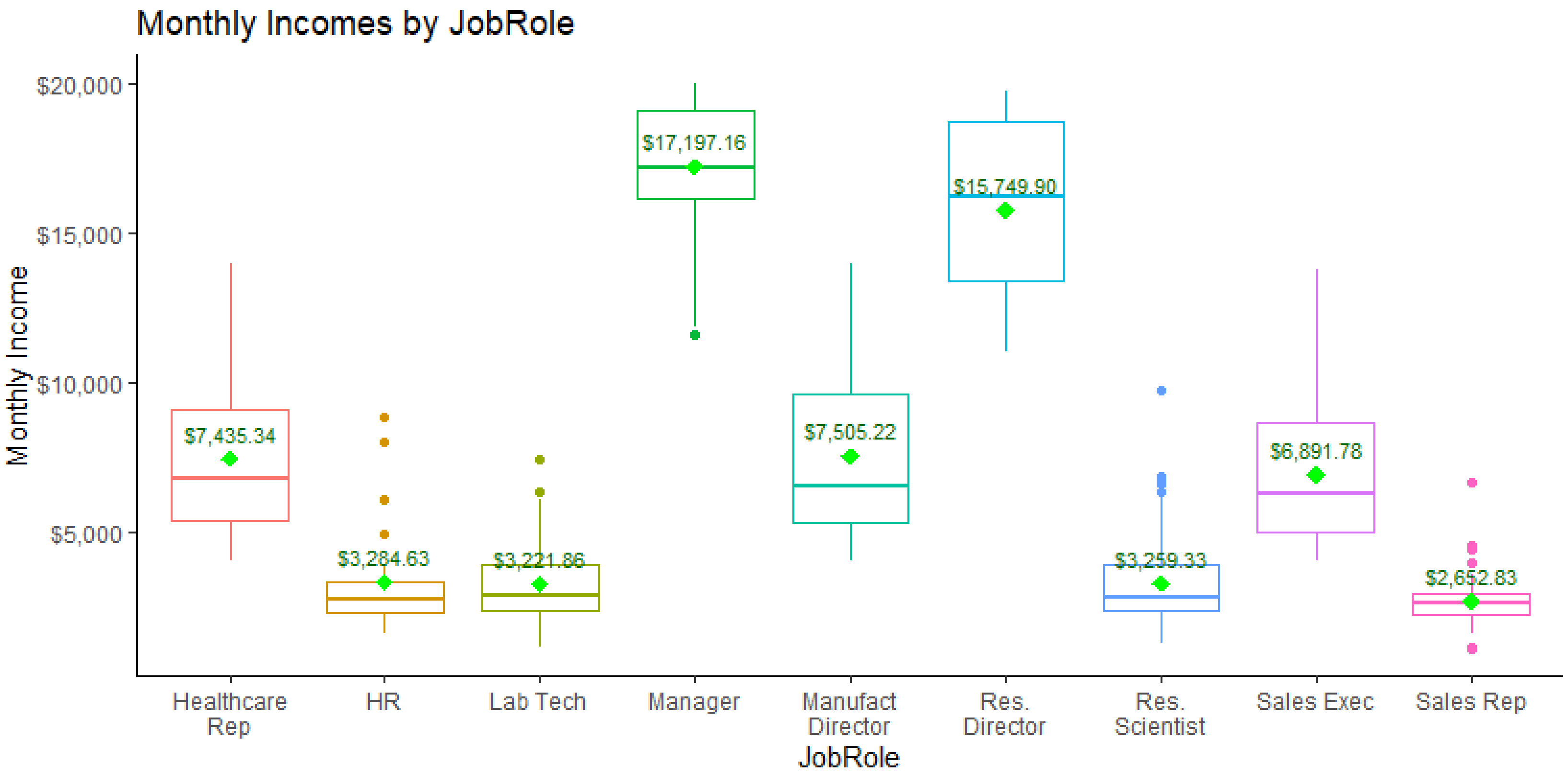


JobRole	OT:Yes Monthly Income	OT:No Monthly Income	Diff
Sales Rep	$2,369.00	$2,798.80	$429.80
Res. Scientist	$3,342.86	$3,208.59	-$134.27
Lab Tech	$3,487.30	$3,148.86	-$338.44
HR	$3,581.33	$3,199.86	-$381.48
Sales Exec	$6,831.90	$6,916.84	$84.94
Healthcare Rep	$7,613.41	$7,362.80	-$250.61
Manufact Director	$7,876.61	$7,371.75	-$504.86
Res. Director	$15,632.31	$15,803.66	$171.34
Manager	$16,617.20	$17,338.61	$721.41

Monthly Income

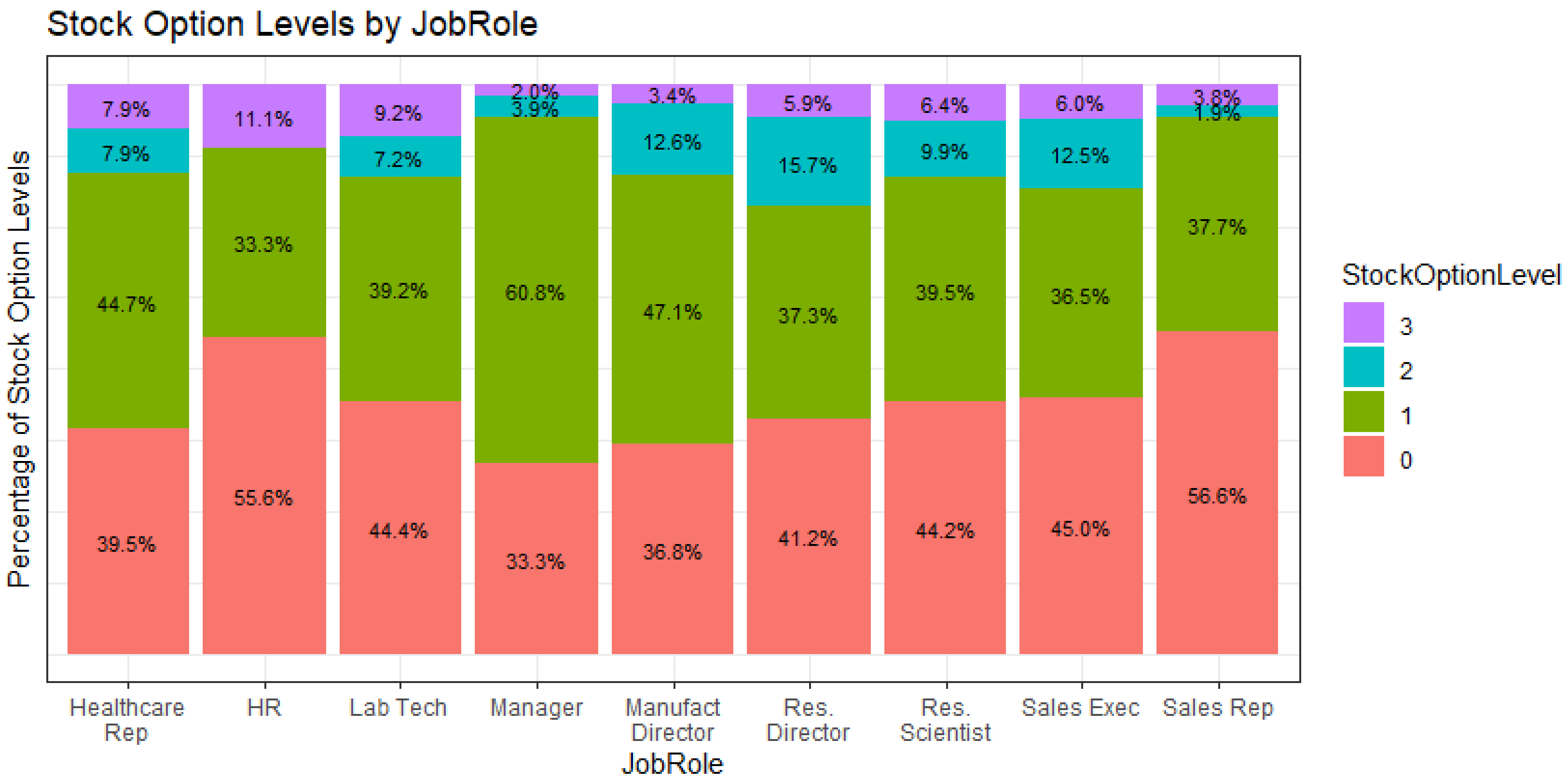
- Mean incomes populated in chart and marked with green diamond

- p-value = 2.41e-07 shows statistical significance
- Lowest independent p-value for all numerical variables
- *T-tests and Chi-Sq Tests in Appendix



Stock Option Levels

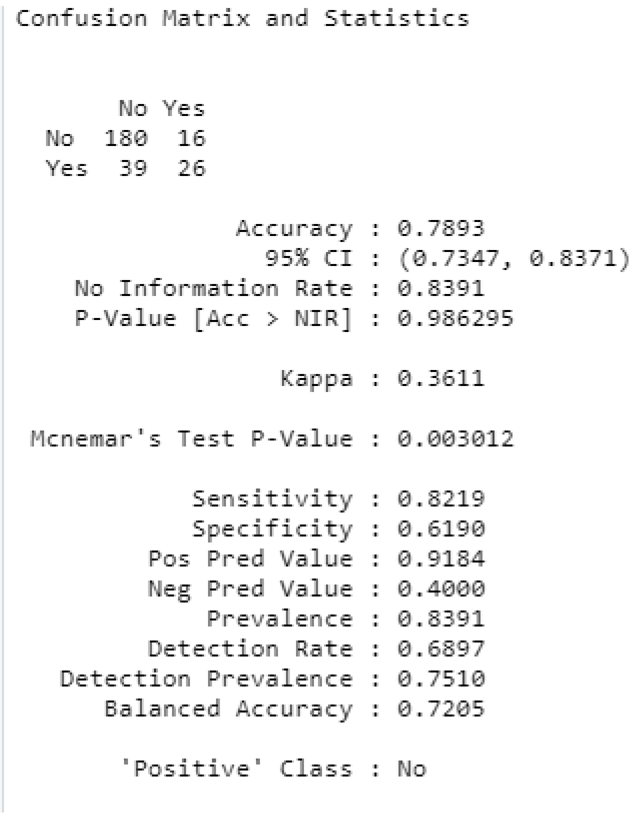
- 43.6% Employees Do Not have Stock Options
- 40.8% = Level 1
- 9.3% = Level 2
- 6.3% = Level 3
- p-value = 3.724e-12
- 2nd Lowest p-value of all categorical variables
- *T-tests and Chi-Sq Tests in Appendix



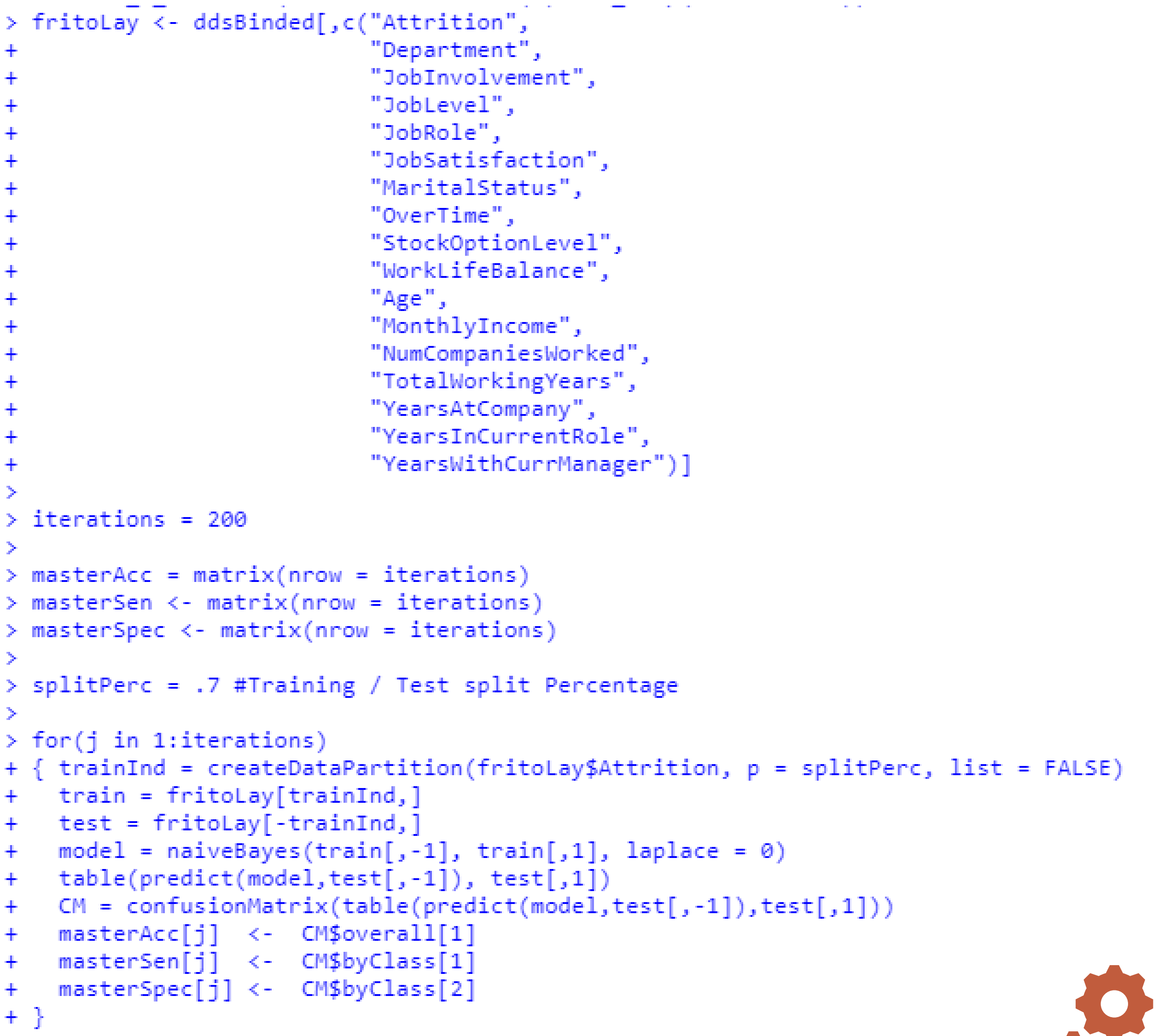
The Top 3 Factors Affecting Attrition

1. OverTime p-value = 2.33e-15 (lowest independent p-value for all categorical variables)
2. MonthlyIncome p-value = 2.41e-07 (lowest independent p-value for all numerical variables)
3. StockOptionLevel p-value = 3.724e-12 (2nd lowest independent p-value for all categorical variables)

Attrition Model



```
Confusion Matrix and Statistics


      No Yes
  No  180  16
  Yes  39  26

               Accuracy : 0.7893
                 95% CI : (0.7347, 0.8371)
    No Information Rate : 0.8391
    P-Value [Acc > NIR] : 0.986295

                  Kappa : 0.3611

 Mcnemar's Test P-Value : 0.003012

            Sensitivity : 0.8219
            Specificity : 0.6190
         Pos Pred Value : 0.9184
         Neg Pred Value : 0.4000
             Prevalence : 0.8391
         Detection Rate : 0.6897
   Detection Prevalence : 0.7510
      Balanced Accuracy : 0.7205

       'Positive' Class : No
```



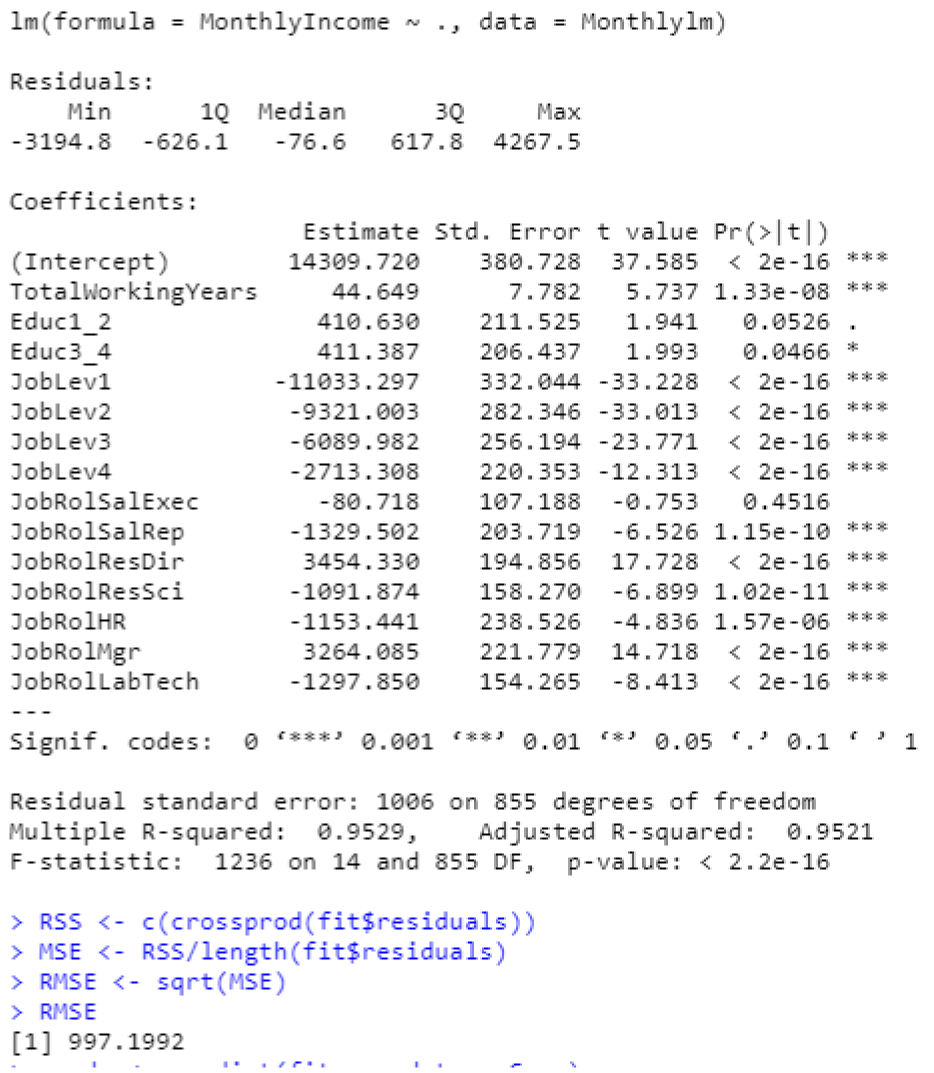
```
> fritoLay <- ddsBinded[,c("Attrition",
+                          "Department",
+                          "JobInvolvement",
+                          "JobLevel",
+                          "JobRole",
+                          "JobSatisfaction",
+                          "MaritalStatus",
+                          "OverTime",
+                          "StockOptionLevel",
+                          "WorkLifeBalance",
+                          "Age",
+                          "MonthlyIncome",
+                          "NumCompaniesWorked",
+                          "TotalWorkingYears",
+                          "YearsAtCompany",
+                          "YearsInCurrentRole",
+                          "YearsWithCurrManager")]
> 
> iterations = 200
> 
> masterAcc = matrix(nrow = iterations)
> masterSen <- matrix(nrow = iterations)
> masterSpec <- matrix(nrow = iterations)
> 
> splitPerc = .7 #Training / Test split Percentage
> 
> for(j in 1:iterations)
+ { trainInd = createDataPartition(fritoLay$Attrition, p = splitPerc, list = FALSE)
+   train = fritoLay[trainInd,]
+   test = fritoLay[-trainInd,]
+   model = naiveBayes(train[,-1], train[,1], laplace = 0)
+   table(predict(model,test[,-1]), test[,1])
+   CM = confusionMatrix(table(predict(model,test[,-1]),test[,1]))
+   masterAcc[j]  <-  CM$overall[1]
+   masterSen[j]  <-  CM$byClass[1]
+   masterSpec[j] <-  CM$byClass[2]
+ }
```

Job Role Trends by Average Scores/ Values

- Age (18-60 years)
  - 47.5 – average oldest - Managers
  - 30.5 – average youngest age – Sales Reps
- Distance From Home (1-29 miles)
  - 9.8 miles – farthest average - Healthcare Reps
  - 7.9 miles – shortest average - Sales Reps
- Environment Satisfaction (1-4)
  - 3.0 – Manufacturing Directors
  - 2.5 – Research Directors
- Job Involvement (1-4)
  - 2.9 – Research Directors
  - 2.6 – Sales Reps
- Job Satisfaction (1-4)
  - 2.8 – Healthcare Reps
  - 2.5 – Research Directors
- Percent Salary Hike
  - 15.7% - Manufacturing Directors
  - 14.9% - Research Directors
  - 15.2% - Company Average
- Training Times Last Year (1-7)
  - 4.2 – Lab Technicians
  - 3.6 – Research Scientists
- Work Life Balance (1-4)
  - 3.0 - HR has best
  - 2.7 - Healthcare Reps
- Sales Reps have the most low Averages by Category
- Managers have the most top averages by Category

Monthly Salary Estimator

- Used Linear Model (lm)
- Model:
  - Monthly Income = $14,309.72 + $44.65*TotalWorkingYears + $410.63*Educ1_2 + $411.39*Educ3_4 - 11033.3*JobLev1 - $9,321*JobLev2 - $6,089.98JobLev3 - $2,713.31*JobLev4 - $80.72*JobRolSalExec - $1,329.5*JobRolSalRep + $3,454.33*JobRolResDir - $1,091.87*JobRolResSci - $1,153.44*JobRolHR + $3,264.08*JobRolMgr - $1,297.85JobRolLabTech
- After 100 Iterations in Cross Validation
- RMSE
  - min: 952.95
  - mean: 992.88
  - max: 1055.74
- R-Squared
  - min: 0.946
  - mean: 0.953
  - max: 0.959



```
lm(formula = MonthlyIncome ~ ., data = Monthlylm)

Residuals:
    Min      1Q  Median      3Q     Max 
-3194.8  -626.1   -76.6   617.8  4267.5 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      14309.720    380.728  37.585  < 2e-16 ***
TotalWorkingYears    44.649      7.782   5.737 1.33e-08 ***
Educ1_2            410.630    211.525   1.941   0.0526 .  
Educ3_4            411.387    206.437   1.993   0.0466 *  
JobLev1         -11033.297    332.044 -33.228  < 2e-16 ***
JobLev2          -9321.003    282.346 -33.013  < 2e-16 ***
JobLev3          -6089.982    256.194 -23.771  < 2e-16 ***
JobLev4          -2713.308    220.353 -12.313  < 2e-16 ***
JobRolSalExec      -80.718    107.188  -0.753   0.4516    
JobRolSalRep     -1329.502    203.719  -6.526 1.15e-10 ***
JobRolResDir      3454.330    194.856  17.728  < 2e-16 ***
JobRolResSci     -1091.874    158.270  -6.899 1.02e-11 ***
JobRolHR         -1153.441    238.526  -4.836 1.57e-06 ***
JobRolMgr         3264.085    221.779  14.718  < 2e-16 ***
JobRolLabTech    -1297.850    154.265  -8.413  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1006 on 855 degrees of freedom
Multiple R-squared:  0.9529,    Adjusted R-squared:  0.9521 
F-statistic:  1236 on 14 and 855 DF,  p-value: < 2.2e-16

> RSS <- c(crossprod(fit$residuals))
> MSE <- RSS/length(fit$residuals)
> RMSE <- sqrt(MSE)
> RMSE
[1] 997.1992
```

Model Interpretation

- Monthly Income = $14,309.72 + $44.65*TotalWorkingYears + $410.63*Educ1_2 + $411.39*Educ3_4 - 11033.3*JobLev1 - $9,321*JobLev2 - $6,089.98JobLev3 - $2,713.31*JobLev4 - $80.72*JobRolSalExec - $1,329.5*JobRolSalRep + $3,454.33*JobRolResDir - $1,091.87*JobRolResSci -$1,153.44*JobRolHR + $3,264.08*JobRolMgr - $1,297.85JobRolLabTech
- Reference Categorical Variables – this means with all the other variables we will discuss below are set to 0, we use the intercept by itself plus Working Years * 44.65
  - Education Level 5
  - Job Level 5
  - Job Role of Healthcare Rep and Manufacturing Director
- Before we start adding anything to the model in terms of inputs, we start with $14,309.72 (intercept) monthly income
- For all other variables held constant, for every company an employee has worked at we multiply that number by $44.65 and add that to the monthly income.
- Categorical Variables
  - If an employee has an Education Level 1 or 2, we add $410.63 to the monthly income
  - If an employee has an Education Level of 3 or 4, we add $411.39 to the monthly income
  - If an employee has Job Level 1, we subtract $11,033.30 from the monthly income
  - If an employee has Job Level 2, we subtract $9,321 from the monthly income
  - If an employee has Job Level 3, we subtract $6,089.98 from the monthly income
  - If an employee has Job Level 4, we subtract $2,713.31 from the monthly income
  - If an employee is a Sales Exec, we subtract $80,72 from the monthly income
  - If an employee is a Sales Rep, we subtract $1,329.50 from the monthly income
  - If an employee is a Research Director, we add $3,454.33 to the monthly income
  - If an employee is a Research Scientist, we subtract $1,091.87 from the monthly income
  - If an employee is in Human Resources, we subtract $1,153.44 from the monthly income
  - If an employee is a Manager, we add $3,264.08 to the monthly income
  - If an employee is a Laboratory Technician, we subtract $1,297.85 from the monthly income

Monthly Income – fun with math!

- Model:

Monthly Income = $14,309.72 + $44.65*TotalWorkingYears + $410.63*Educ1_2 + $411.39*Educ3_4 - 11033.3*JobLev1 - $9,321*JobLev2 - $6,089.98JobLev3 - $2,713.31*JobLev4 - $80.72*JobRolSalExec - $1,329.5*JobRolSalRep + $3,454.33*JobRolResDir - $1,091.87*JobRolResSci -$1,153.44*JobRolHR + $3,264.08*JobRolMgr - $1,297.85JobRolLabTech

- Job Role: Account Executive
- Education Level: 2
- Job Level: 2
- Years Work Experience: 2
- Est Monthly Income = $14,309.72 + $44.65*2yrs + $410.63*1 (education level) - $9,321.00*1 (JobLevel2) - $80.72*1 (Sales Exec)
- Est Monthly Income = $14,309.72 + $89.30 - $410.63 - $9,321.00 - $80.72
- Est Monthly Income = $14,309.72 - $9,723.05
- Est Monthly Income = $4,586.67 gross

Key Takeaways

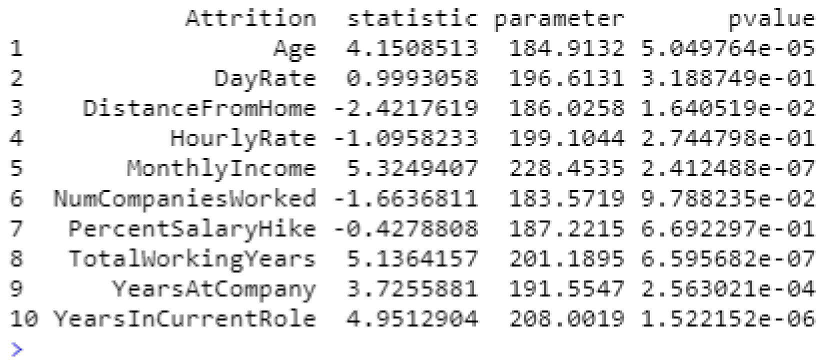
- Overtime, Monthly Income and Stock Options are very important when predicting employee Attrition
- 71% of your employees don’t earn overtime, but we don’t have any idea of bonuses or other perks
- People making OT earn on average $255.98 less per month than people who do not earn OT
- Monthly Income is the most statistically significant continuous variable we have in determining Attrition
- 43.6% of employees do not have stock options and that is a top 3 contributing factor to Attrition
- Sales Reps seem to have the lowest average scores in term of categorical variables, but they are also the youngest on average with the least amount of experience
- Manager seem to have the most high scores and are the oldest on average with the most experience
- Total Working Years, Education, Job Level and Job Role show the most statistical significance for estimating monthly income

Appendix

Attrition

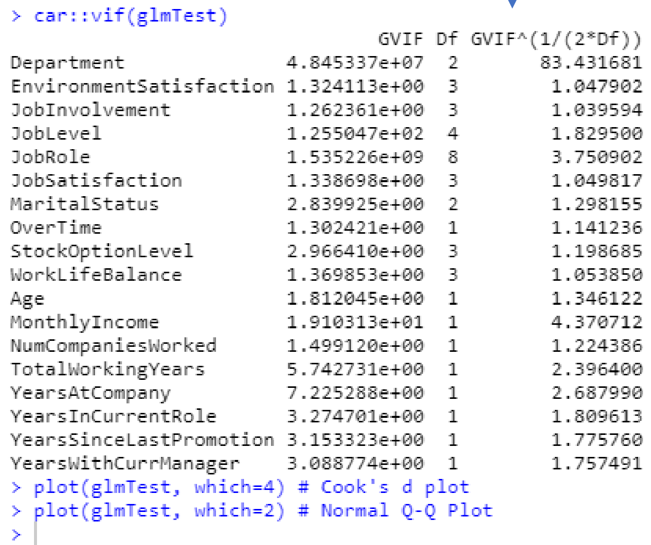
- Double Checking the Boruta classifier Using GLM
- Lots of High P-values, but very low VIFs
- After trying to tune it better, found that the algorithm is pretty good at its job

Independent

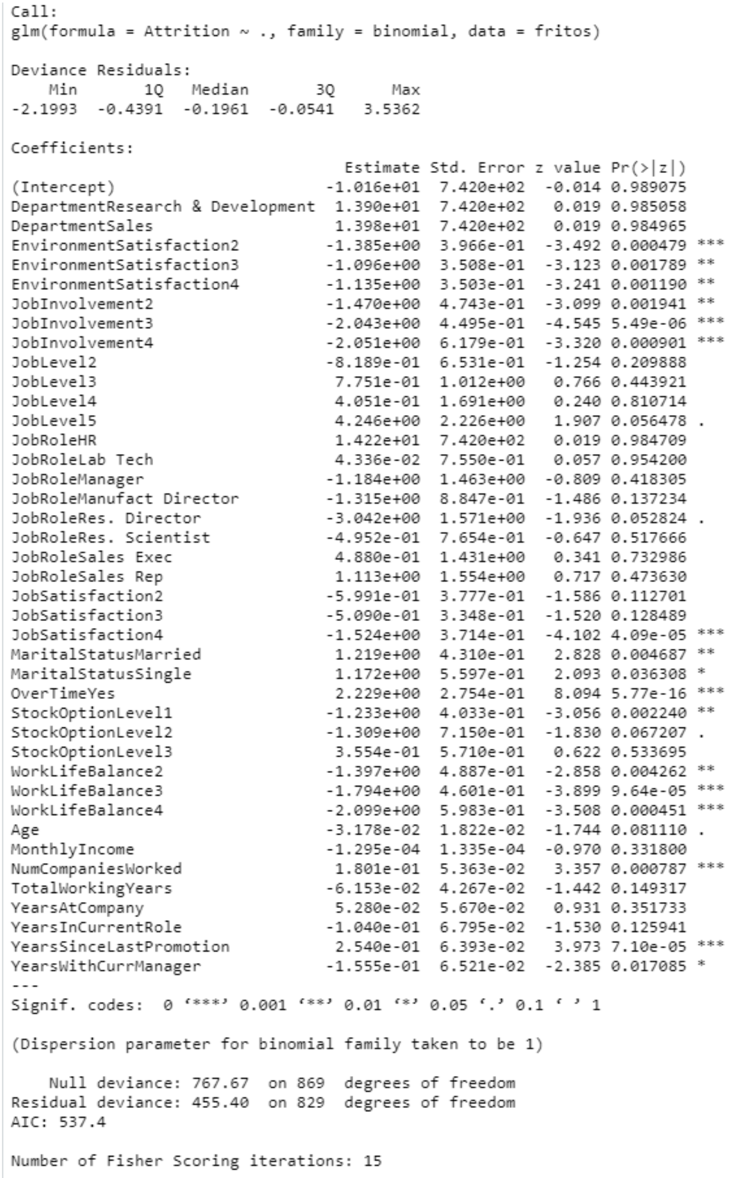


```
          Attrition  statistic parameter       pvalue
1               Age  4.1508513  184.9132 5.049764e-05
2           DayRate  0.9993058  196.6131 3.188749e-01
3  DistanceFromHome -2.4217619  186.0258 1.640519e-02
4        HourlyRate -1.0958233  199.1044 2.744798e-01
5     MonthlyIncome  5.3249407  228.4535 2.412488e-07
6 NumCompaniesWorked -1.6636811  183.5719 9.788235e-02
7   PercentSalaryHike -0.4278808  187.2215 6.692297e-01
8   TotalWorkingYears  5.1364157  201.1895 6.595682e-07
9      YearsAtCompany  3.7255881  191.5547 2.563021e-04
10 YearsInCurrentRole  4.9512904  208.0019 1.522152e-06
>
```

Grouped



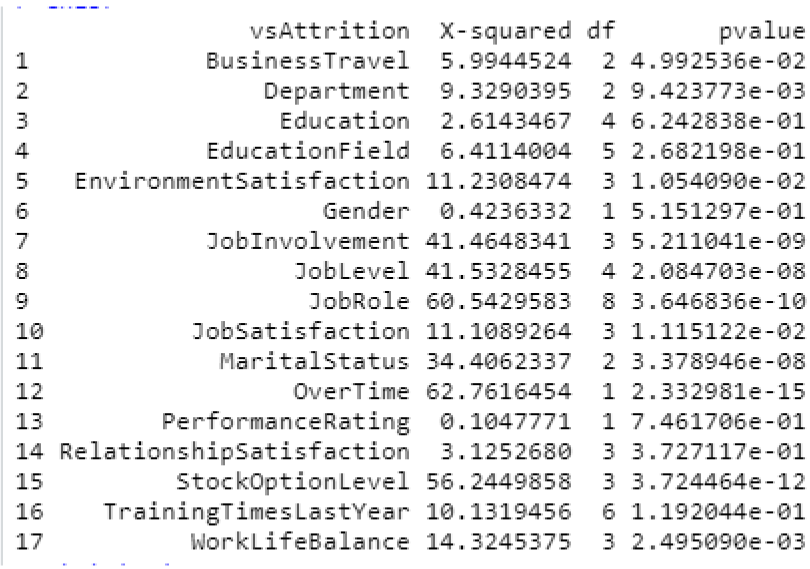
```
> car::vif(glmTest)
                                GVIF Df GVIF^(1/(2*Df))
Department              4.845337e+07  2       83.431681
EnvironmentSatisfaction 1.324113e+00  3        1.047902
JobInvolvement          1.262361e+00  3        1.039594
JobLevel                1.255047e+02  4        1.829500
JobRole                 1.535226e+09  8        3.750902
JobSatisfaction         1.338698e+00  3        1.049817
MaritalStatus           2.839925e+00  2        1.298155
OverTime                1.302421e+00  1        1.141236
StockOptionLevel        2.966410e+00  3        1.198685
WorkLifeBalance         1.369853e+00  3        1.053850
Age                     1.812045e+00  1        1.346122
MonthlyIncome           1.910313e+01  1        4.370712
NumCompaniesWorked      1.499120e+00  1        1.224386
TotalWorkingYears       5.742731e+00  1        2.396400
YearsAtCompany          7.225288e+00  1        2.687990
YearsInCurrentRole      3.274701e+00  1        1.809613
YearsSinceLastPromotion 3.153323e+00  1        1.775760
YearsWithCurrManager    3.088774e+00  1        1.757491
> plot(glmTest, which=4) # Cook's d plot
> plot(glmTest, which=2) # Normal Q-Q Plot
> 
```



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

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.1993  -0.4391  -0.1961  -0.0541   3.5362  

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)    
(Intercept)                     -1.016e+01  7.420e+02  -0.014 0.989075    
DepartmentResearch & Development  1.390e+01  7.420e+02   0.019 0.985058    
DepartmentSales                  1.398e+01  7.420e+02   0.019 0.984965    
EnvironmentSatisfaction2        -1.385e+00  3.966e-01  -3.492 0.000479 ***
EnvironmentSatisfaction3        -1.096e+00  3.508e-01  -3.123 0.001789 ** 
EnvironmentSatisfaction4        -1.135e+00  3.503e-01  -3.241 0.001190 ** 
JobInvolvement2                 -1.470e+00  4.743e-01  -3.099 0.001941 ** 
JobInvolvement3                 -2.043e+00  4.495e-01  -4.545 5.49e-06 ***
JobInvolvement4                 -2.051e+00  6.179e-01  -3.320 0.000901 ***
JobLevel2                       -8.189e-01  6.531e-01  -1.254 0.209888    
JobLevel3                        7.751e-01  1.012e+00   0.766 0.443921    
JobLevel4                        4.051e-01  1.691e+00   0.240 0.810714    
JobLevel5                        4.246e+00  2.226e+00   1.907 0.056478 .  
JobRoleHR                        1.422e+01  7.420e+02   0.019 0.984709    
JobRoleLab Tech                  4.336e-02  7.550e-01   0.057 0.954200    
JobRoleManager                  -1.184e+00  1.463e+00  -0.809 0.418305    
JobRoleManufact Director        -1.315e+00  8.847e-01  -1.486 0.137234    
JobRoleRes. Director            -3.042e+00  1.571e+00  -1.936 0.052824 .  
JobRoleRes. Scientist           -4.952e-01  7.654e-01  -0.647 0.517666    
JobRoleSales Exec                4.880e-01  1.431e+00   0.341 0.732986    
JobRoleSales Rep                 1.113e+00  1.554e+00   0.717 0.473630    
JobSatisfaction2                -5.991e-01  3.777e-01  -1.586 0.112701    
JobSatisfaction3                -5.090e-01  3.348e-01  -1.520 0.128489    
JobSatisfaction4                -1.524e+00  3.714e-01  -4.102 4.09e-05 ***
MaritalStatusMarried             1.219e+00  4.310e-01   2.828 0.004687 ** 
MaritalStatusSingle              1.172e+00  5.597e-01   2.093 0.036308 *  
OverTimeYes                      2.229e+00  2.754e-01   8.094 5.77e-16 ***
StockOptionLevel1               -1.233e+00  4.033e-01  -3.056 0.002240 ** 
StockOptionLevel2               -1.309e+00  7.150e-01  -1.830 0.067207 .  
StockOptionLevel3                3.554e-01  5.710e-01   0.622 0.533695    
WorkLifeBalance2                -1.397e+00  4.887e-01  -2.858 0.004262 ** 
WorkLifeBalance3                -1.794e+00  4.601e-01  -3.899 9.64e-05 ***
WorkLifeBalance4                -2.099e+00  5.983e-01  -3.508 0.000451 ***
Age                             -3.178e-02  1.822e-02  -1.744 0.081110 .  
MonthlyIncome                   -1.295e-04  1.335e-04  -0.970 0.331800    
NumCompaniesWorked               1.801e-01  5.363e-02   3.357 0.000787 ***
TotalWorkingYears               -6.153e-02  4.267e-02  -1.442 0.149317    
YearsAtCompany                   5.280e-02  5.670e-02   0.931 0.351733    
YearsInCurrentRole              -1.040e-01  6.795e-02  -1.530 0.125941    
YearsSinceLastPromotion          2.540e-01  6.393e-02   3.973 7.10e-05 ***
YearsWithCurrManager            -1.555e-01  6.521e-02  -2.385 0.017085 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 767.67  on 869  degrees of freedom
Residual deviance: 455.40  on 829  degrees of freedom
AIC: 537.4

Number of Fisher Scoring iterations: 15
```

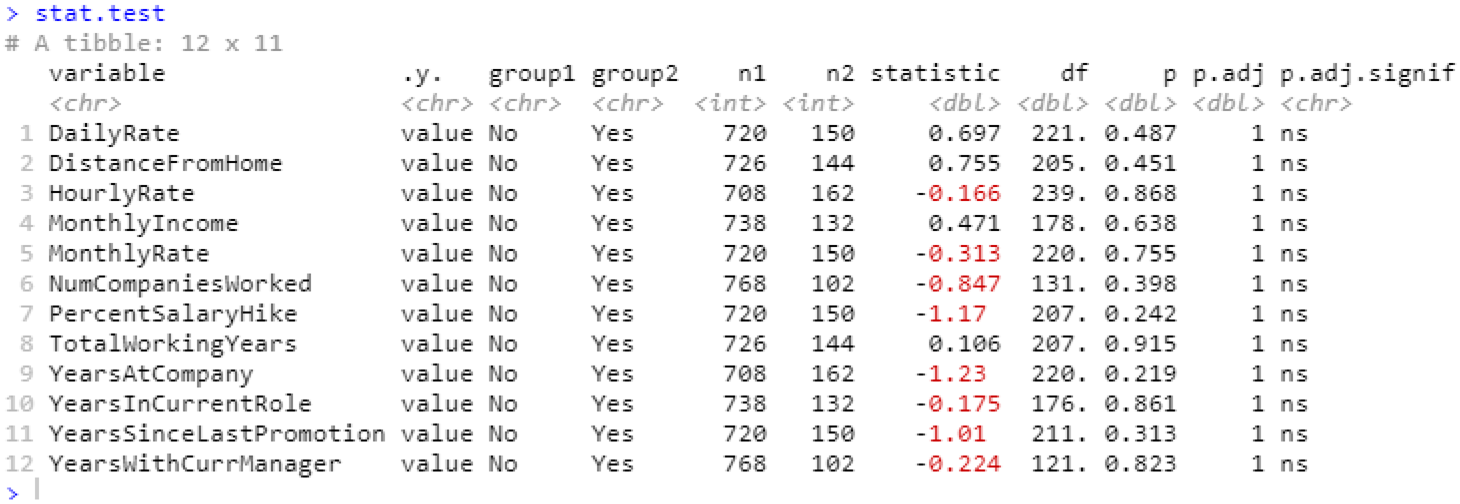
ChiSquared Tests

Independent vs Attrition



```
                vsAttrition  X-squared df       pvalue
1            BusinessTravel  5.9944524  2 4.992536e-02
2                Department  9.3290395  2 9.423773e-03
3                 Education  2.6143467  4 6.242838e-01
4            EducationField  6.4114004  5 2.682198e-01
5   EnvironmentSatisfaction 11.2308474  3 1.054090e-02
6                    Gender  0.4236332  1 5.151297e-01
7            JobInvolvement 41.4648341  3 5.211041e-09
8                  JobLevel 41.5328455  4 2.084703e-08
9                   JobRole 60.5429583  8 3.646836e-10
10          JobSatisfaction 11.1089264  3 1.115122e-02
11            MaritalStatus 34.4062337  2 3.378946e-08
12                 OverTime 62.7616454  1 2.332981e-15
13        PerformanceRating  0.1047771  1 7.461706e-01
14 RelationshipSatisfaction  3.1252680  3 3.727117e-01
15         StockOptionLevel 56.2449858  3 3.724464e-12
16    TrainingTimesLastYear 10.1319456  6 1.192044e-01
17          WorkLifeBalance 14.3245375  3 2.495090e-03
```

Group Test vs Attrition



```
> stat.test
# A tibble: 12 x 11
   variable                .y.   group1 group2    n1    n2 statistic    df     p p.adj p.adj.signif
   <chr>                   <chr> <chr>  <chr>  <int> <int>     <dbl> <dbl> <dbl> <dbl> <chr>
 1 DailyRate               value No     Yes      720   150     0.697  221. 0.487     1 ns
 2 DistanceFromHome        value No     Yes      726   144     0.755  205. 0.451     1 ns
 3 HourlyRate              value No     Yes      708   162    -0.166  239. 0.868     1 ns
 4 MonthlyIncome           value No     Yes      738   132     0.471  178. 0.638     1 ns
 5 MonthlyRate             value No     Yes      720   150    -0.313  220. 0.755     1 ns
 6 NumCompaniesWorked      value No     Yes      768   102    -0.847  131. 0.398     1 ns
 7 PercentSalaryHike       value No     Yes      720   150    -1.17   207. 0.242     1 ns
 8 TotalWorkingYears       value No     Yes      726   144     0.106  207. 0.915     1 ns
 9 YearsAtCompany          value No     Yes      708   162    -1.23   220. 0.219     1 ns
10 YearsInCurrentRole      value No     Yes      738   132    -0.175  176. 0.861     1 ns
11 YearsSinceLastPromotion value No     Yes      720   150    -1.01   211. 0.313     1 ns
12 YearsWithCurrManager    value No     Yes      768   102    -0.224  121. 0.823     1 ns
>
```