Brett Bertucio 330 Project

2024-06-05

“Who Will Go and Who Will Stay?”

Purpose

As I’ve explored my own career path in data analytics, I’ve become interested in the growing field of people analytics. This project mines data from an synthetic HR dataset (https://www.kaggle.com/datasets/rhuebner/human-resources-data-set/data) created by Drs. Carla Patalano and Rich Huebner for use at the New England Institute of Business.

The dataset contains 36 variables for 300+ employees ranging from their date of birth to salary to reasons for voluntarily leaving the company. I am most interested in the latter. The purpose of this project is to model and predict employee retention. The construction of the model should reveal which factors most lead to retention and which factors lead to voluntary termination. Models can also be used to predict whether particular employees will leave.

Data Sources, Clearning, and Creating New Derived Variables

Because I’m particularly interested in employees who choose to leave, I’ve created individual binary variables for different employee statuses (Active, Terminated for Cause, Voluntarily Terminated).

```
library(tidyverse)
library(lubridate)

url = "https://drive.google.com/uc?export=download&id=1sjCnxJAICCfW3tU3QzzFRwBLr8CPFL2C"
hrdata = read.csv(url)
str(hrdata)
```

```
## 'data.frame':    311 obs. of  36 variables:
##  $ Employee_Name             : chr  "Adinolfi, Wilson  K" "Ait Sidi, Karthikeyan   " "Akinkuo
lie, Sarah" "Alagbe,Trina" ...
##  $ EmpID                     : int  10026 10084 10196 10088 10069 10002 10194 10062 10114 102
50 ...
##  $ MarriedID                 : int  0 1 1 1 0 0 0 0 0 0 ...
##  $ MaritalStatusID           : int  0 1 1 1 2 0 0 4 0 2 ...
##  $ GenderID                  : int  1 1 0 0 0 0 0 1 0 1 ...
##  $ EmpStatusID               : int  1 5 5 1 5 1 1 1 3 1 ...
##  $ DeptID                    : int  5 3 5 5 5 5 4 5 5 3 ...
##  $ PerfScoreID               : int  4 3 3 3 3 4 3 3 3 3 ...
##  $ FromDiversityJobFairID    : int  0 0 0 0 0 0 0 0 1 0 ...
##  $ Salary                    : int  62506 104437 64955 64991 50825 57568 95660 59365 47837 50
178 ...
##  $ Termd                     : int  0 1 1 0 1 0 0 0 0 0 ...
##  $ PositionID                : int  19 27 20 19 19 19 24 19 19 14 ...
##  $ Position                  : chr  "Production Technician I" "Sr. DBA" "Production Technicia
n II" "Production Technician I" ...
##  $ State                     : chr  "MA" "MA" "MA" "MA" ...
##  $ Zip                       : int  1960 2148 1810 1886 2169 1844 2110 2199 1902 1886 ...
##  $ DOB                       : chr  "7/10/1983" "5/5/1975" "9/19/1988" "9/27/1988" ...
##  $ Sex                       : chr  "M " "M " "F" "F" ...
##  $ MaritalDesc               : chr  "Single" "Married" "Married" "Married" ...
##  $ CitizenDesc               : chr  "US Citizen" "US Citizen" "US Citizen" "US Citizen" ...
##  $ HispanicLatino            : chr  "No" "No" "No" "No" ...
##  $ RaceDesc                  : chr  "White" "White" "White" "White" ...
##  $ DateofHire                : chr  "7/5/2011" "3/30/2015" "7/5/2011" "1/7/2008" ...
##  $ DateofTermination         : chr  "" "6/16/2016" "9/24/2012" "" ...
##  $ TermReason                : chr  "N/A-StillEmployed" "career change" "hours" "N/A-StillEmp
loyed" ...
##  $ EmploymentStatus          : chr  "Active" "Voluntarily Terminated" "Voluntarily Terminate
d" "Active" ...
##  $ Department                : chr  "Production       " "IT/IS" "Production       " "Producti
on       " ...
##  $ ManagerName               : chr  "Michael Albert" "Simon Roup" "Kissy Sullivan" "Elijiah G
ray" ...
##  $ ManagerID                 : int  22 4 20 16 39 11 10 19 12 7 ...
##  $ RecruitmentSource         : chr  "LinkedIn" "Indeed" "LinkedIn" "Indeed" ...
##  $ PerformanceScore          : chr  "Exceeds" "Fully Meets" "Fully Meets" "Fully Meets" ...
##  $ EngagementSurvey          : num  4.6 4.96 3.02 4.84 5 5 3.04 5 4.46 5 ...
##  $ EmpSatisfaction           : int  5 3 3 5 4 5 3 4 3 5 ...
##  $ SpecialProjectsCount      : int  0 6 0 0 0 0 4 0 0 6 ...
##  $ LastPerformanceReview_Date: chr  "1/17/2019" "2/24/2016" "5/15/2012" "1/3/2019" ...
##  $ DaysLateLast30            : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Absences                  : int  1 17 3 15 2 15 19 19 4 16 ...
```

```
hrdata$Sex = factor(hrdata$Sex)
hrdata = hrdata %>% mutate_if(is.character, as.factor)
hrdata$DOB = as.Date(hrdata$DOB)
hrdata$DateofHire = as.Date(hrdata$DateofHire)
hrdata$DateofTermination = as.integer(hrdata$DateofTermination)
hrdata$DateofTermination = as.Date(hrdata$DateofTermination)


#New Derived Variables
hrdata = hrdata %>%
  mutate(
    Active = ifelse(EmploymentStatus == "Active", 1, 0),
    TerminatedForCause = ifelse(EmploymentStatus == "Terminated for Cause", 1, 0),
    VoluntarilyTerminated = ifelse(EmploymentStatus == "Voluntarily Terminated", 1, 0)
  )
```

Data Exploration and Further Variable Creation

I use a combination of plots and descriptive statistics to understand the contours of my data. To prepare data for use in a decision tree, in a few cases I've created new binary variables that indicate if a row entry is in a variable value of particular interest. Here, I've created variables that express whether an employee lived in the headquarters state ("InState") and whether the employee was in one of the two most common positions to depart - Production Technitian I or II ("ProductionTech")

```
count(hrdata, TermReason)
```

```
##                          TermReason   n
## 1                  Another position  20
## 2                        attendance   7
## 3                     career change   9
## 4                  Fatal attraction   1
## 5                  gross misconduct   1
## 6                             hours   8
## 7     Learned that he is a gangster   1
## 8  maternity leave - did not return   3
## 9                    medical issues   3
## 10                         military   4
## 11                       more money  11
## 12                N/A-StillEmployed 207
## 13                 no-call, no-show   4
## 14                      performance   4
## 15           relocation out of area   5
## 16                          retiring   4
## 17                 return to school   5
## 18                           unhappy  14
```

```
count(hrdata, CitizenDesc)
```

```
##           CitizenDesc   n
## 1 Eligible NonCitizen  12
## 2         Non-Citizen   4
## 3          US Citizen 295
```

```
count(hrdata, Sex)
```

```
##   Sex   n
## 1   F 176
## 2  M  135
```

```
count(hrdata, FromDiversityJobFairID)
```

```
##   FromDiversityJobFairID   n
## 1                      0 282
## 2                      1  29
```

```
count(hrdata, Position)
```

```
##                       Position   n
## 1                  Accountant I   3
## 2      Administrative Assistant   3
## 3             Area Sales Manager  27
## 4                   BI Developer   4
## 5                    BI Director   1
## 6                            CIO   1
## 7                   Data Analyst   7
## 8                  Data Analyst    1
## 9                 Data Architect   1
## 10       Database Administrator   5
## 11       Director of Operations   1
## 12            Director of Sales   1
## 13         Enterprise Architect   1
## 14                   IT Director   1
## 15               IT Manager - DB   2
## 16            IT Manager - Infra   1
## 17         IT Manager - Support   1
## 18                    IT Support   8
## 19              Network Engineer   5
## 20               President & CEO   1
## 21     Principal Data Architect   1
## 22            Production Manager  14
## 23      Production Technician I 137
## 24     Production Technician II  57
## 25                 Sales Manager   3
## 26           Senior BI Developer   3
## 27      Shared Services Manager   1
## 28             Software Engineer  10
## 29 Software Engineering Manager   1
## 30                Sr. Accountant   2
## 31                       Sr. DBA   2
## 32         Sr. Network Engineer   5
```

```
count(hrdata, MaritalDesc)
```

```
##   MaritalDesc   n
## 1    Divorced  30
## 2     Married 124
## 3   Separated  12
## 4      Single 137
## 5     Widowed   8
```

```
count(hrdata %>% filter(VoluntarilyTerminated == 1), Position)
```

```
##                    Position  n
## 1  Administrative Assistant  1
## 2         Area Sales Manager  2
## 3               Data Analyst  1
## 4              Data Analyst   1
## 5            IT Manager - DB  1
## 6           Network Engineer  1
## 7  Principal Data Architect  1
## 8         Production Manager  4
## 9   Production Technician I 45
## 10 Production Technician II 26
## 11             Sales Manager  1
## 12         Software Engineer  3
## 13                   Sr. DBA  1
```

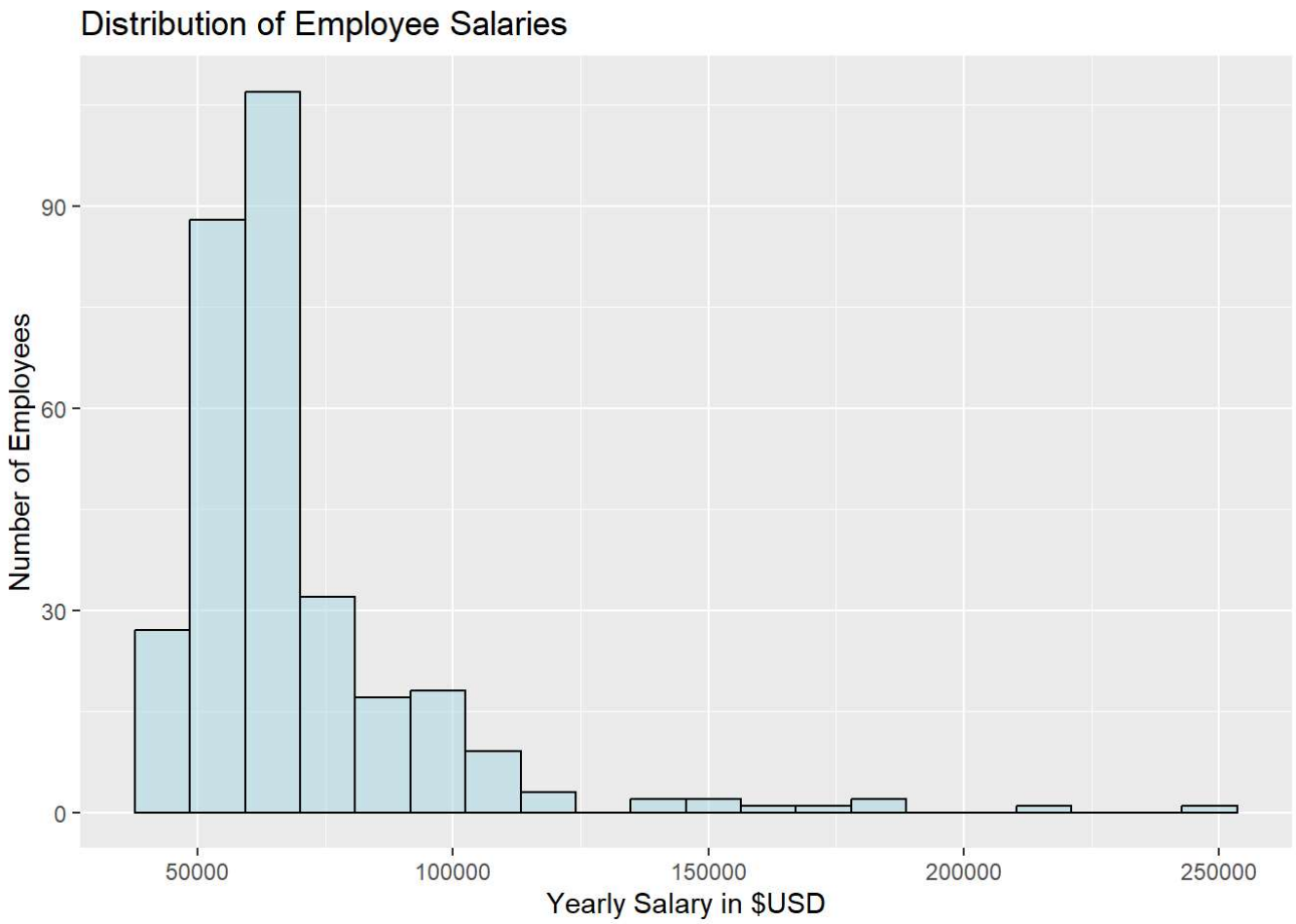
```
count(hrdata, State)
```

```
##    State   n
## 1     AL   1
## 2     AZ   1
## 3     CA   1
## 4     CO   1
## 5     CT   6
## 6     FL   1
## 7     GA   1
## 8     ID   1
## 9     IN   1
## 10    KY   1
## 11    MA 276
## 12    ME   1
## 13    MT   1
## 14    NC   1
## 15    ND   1
## 16    NH   1
## 17    NV   1
## 18    NY   1
## 19    OH   1
## 20    OR   1
## 21    PA   1
## 22    RI   1
## 23    TN   1
## 24    TX   3
## 25    UT   1
## 26    VA   1
## 27    VT   2
## 28    WA   1
```

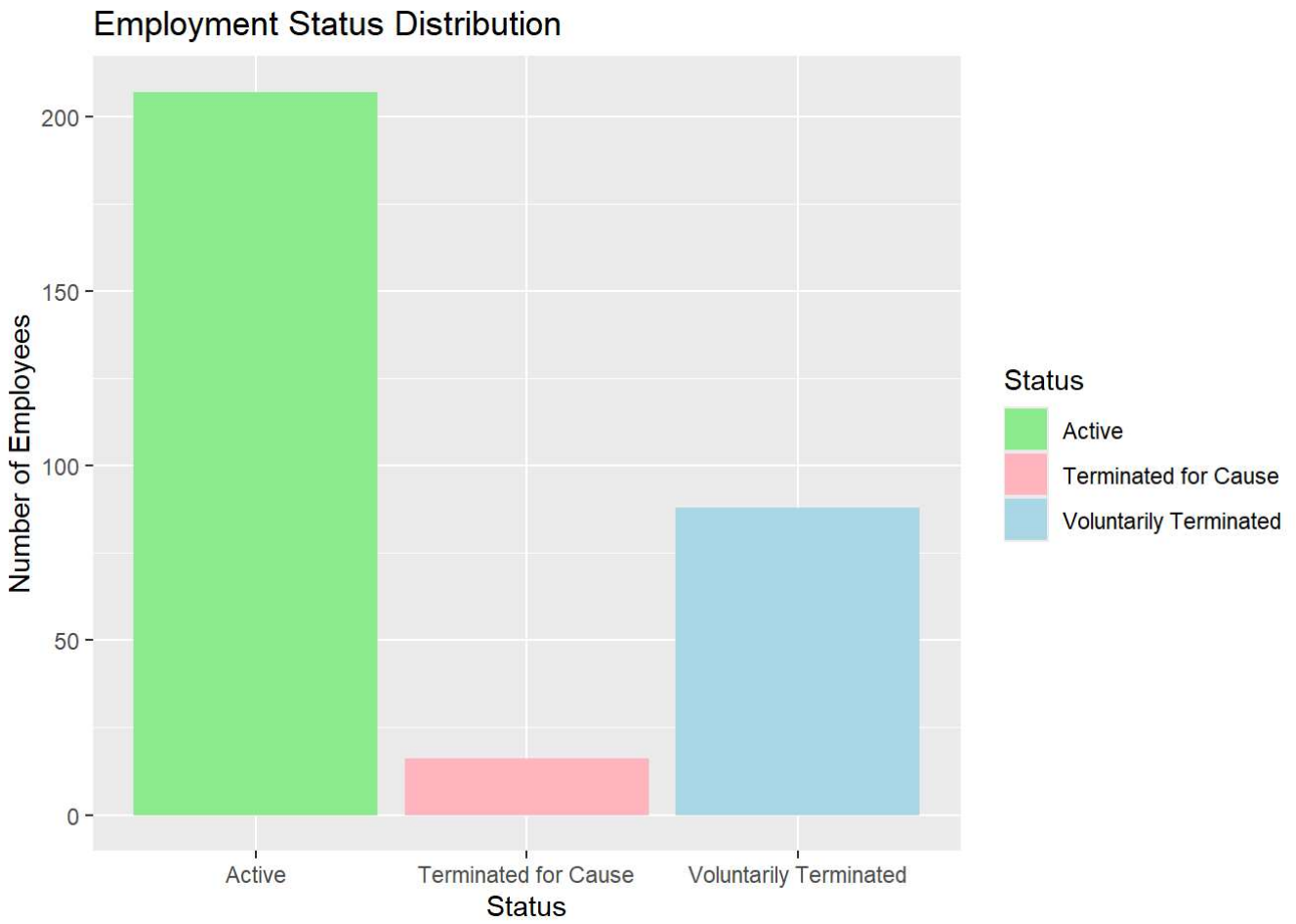
```
hrdata = hrdata %>% mutate(InState = ifelse(State == "MA", 1, 0))
hrdata$ProductionTech = ifelse(hrdata$PositionID %in% c("19", "20"), 1, 0)
```

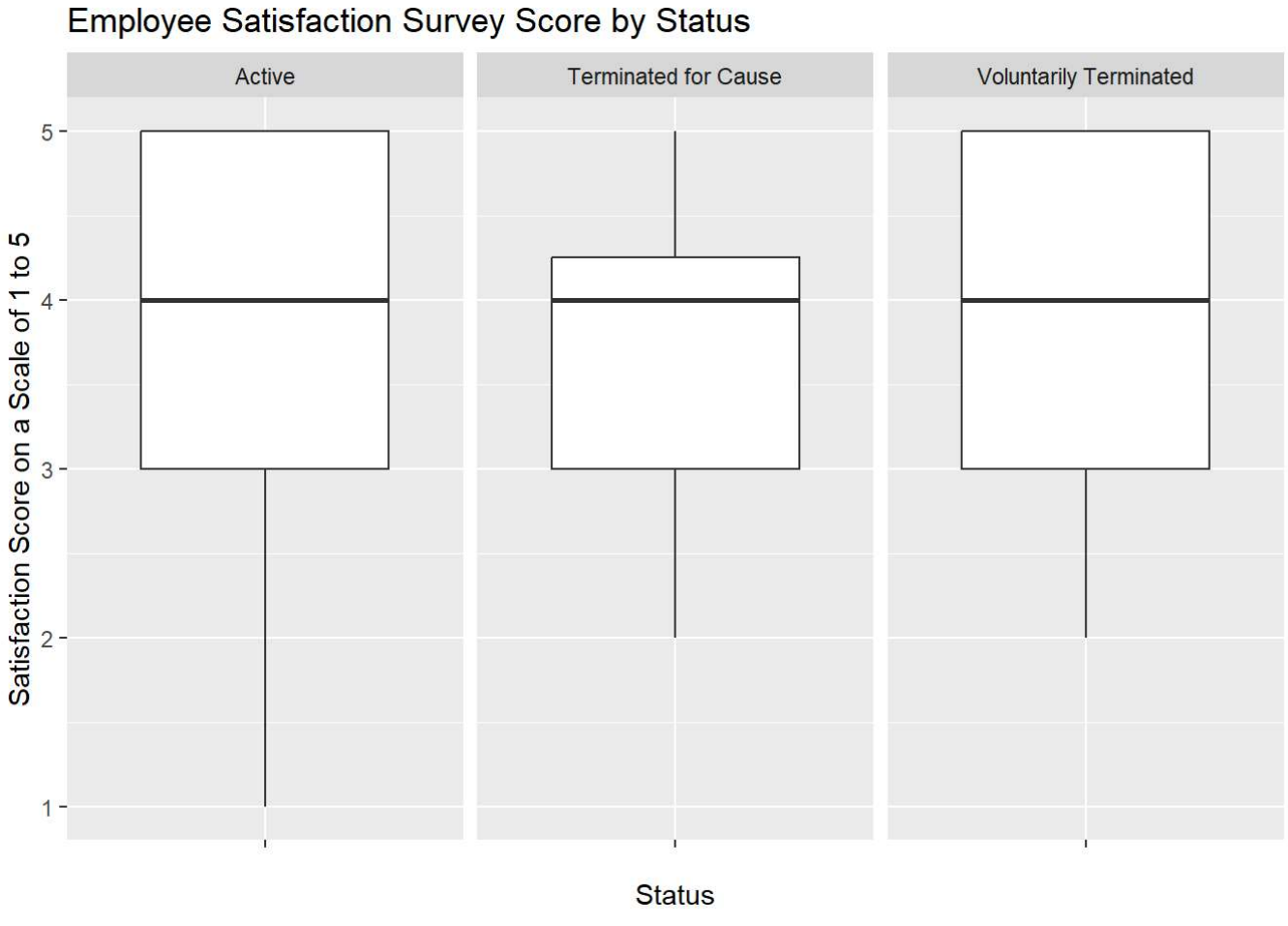
Plotting Data

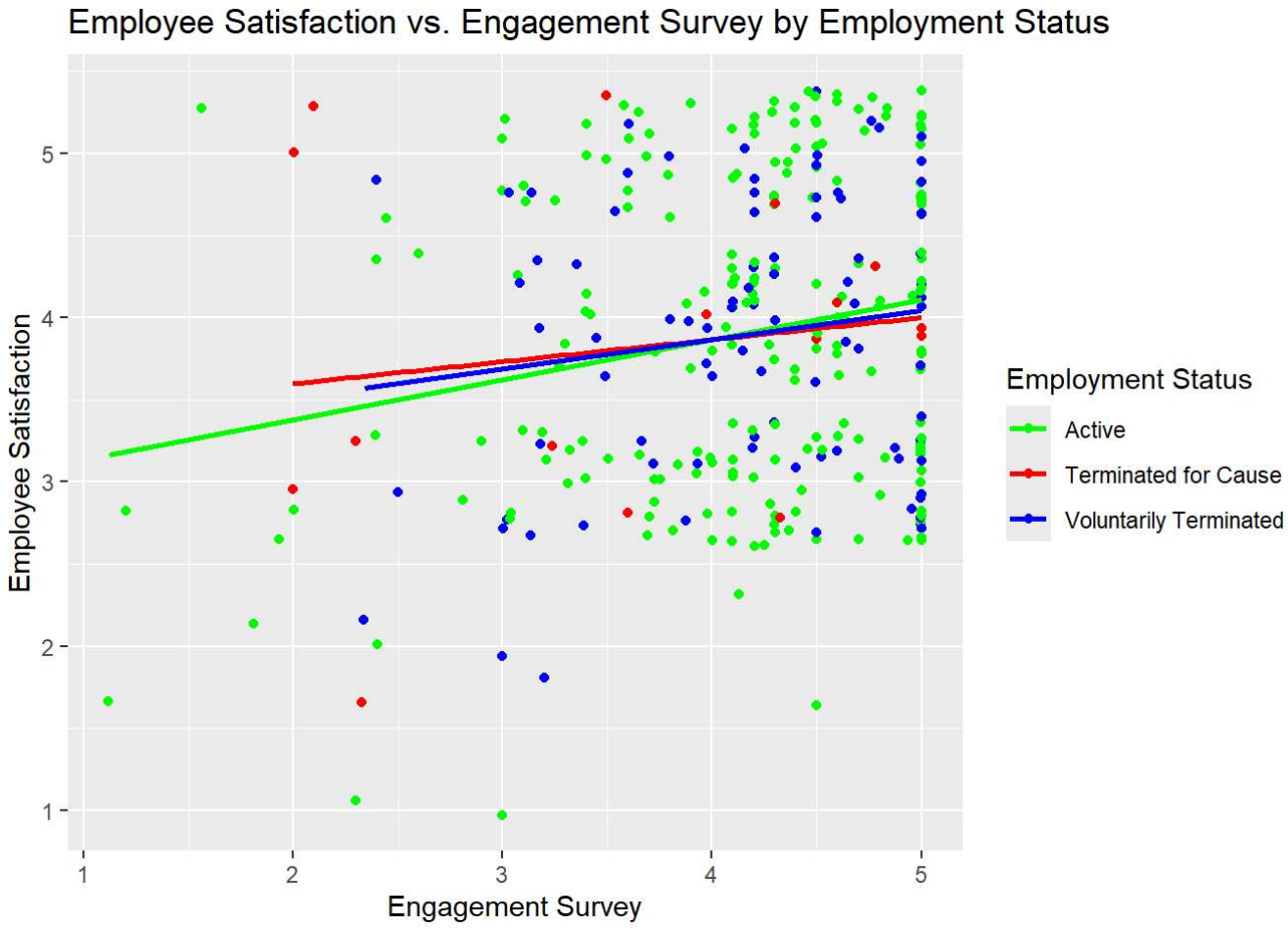
These plots give some context to overall distribution of data and relationships between data. The final plot, a correlation matrix, will be used for feature selection for a logisitc regression model.


Distribution of Employee Salaries
Number of Employees
0
30
60
90
50000
100000
150000
200000
250000
Yearly Salary in $USD

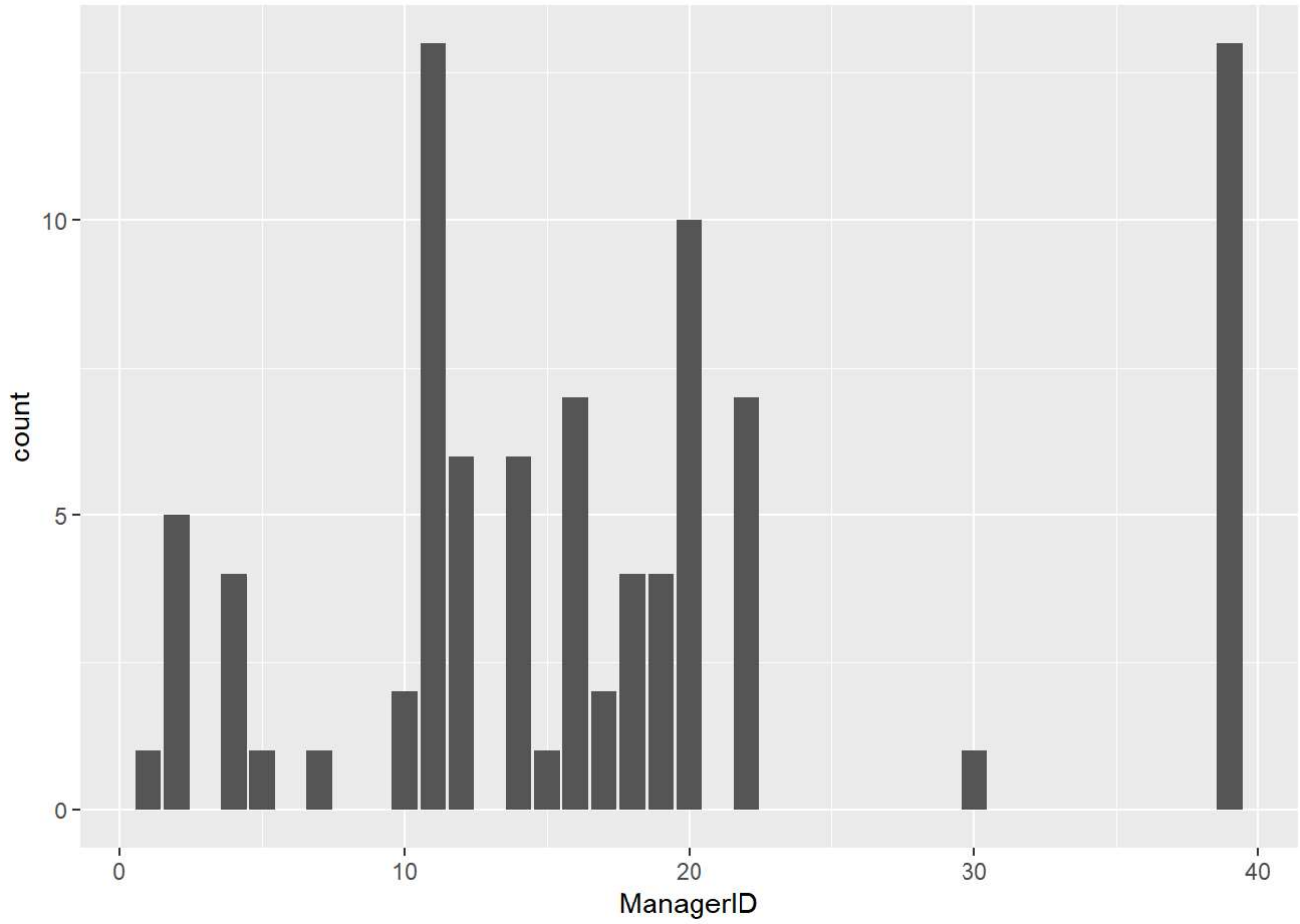

Employee Salary by Status
Active
Terminated for Cause
Voluntarily Terminated
Salary in $USD
50000
100000
150000
200000
250000
Status

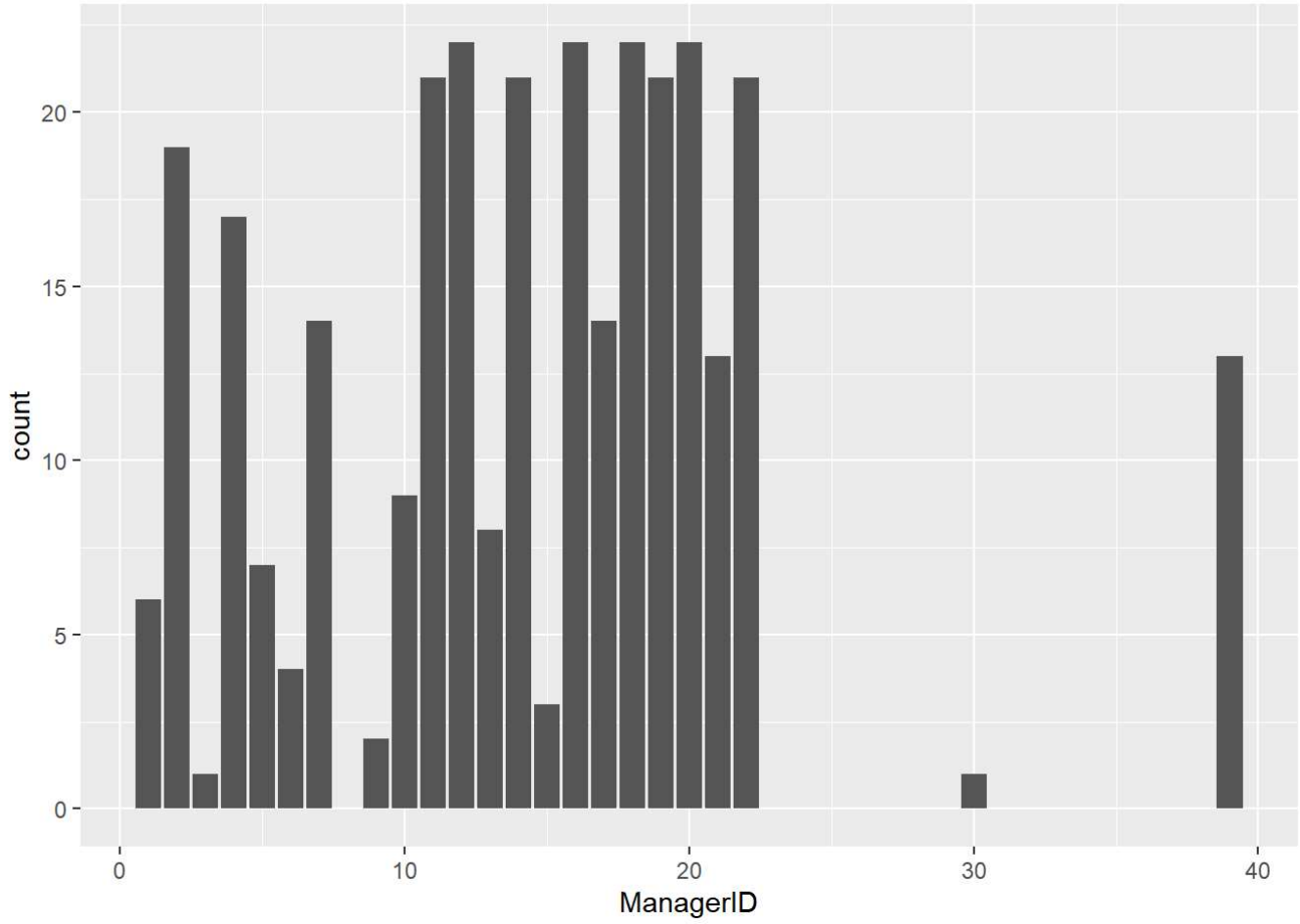

Employment Status Distribution
Number of Employees
200
150
100
50
0
Active
Terminated for Cause
Voluntarily Terminated
Status
Status
Active
Terminated for Cause
Voluntarily Terminated


Employee Satisfaction Survey Score by Status
Active
Terminated for Cause
Voluntarily Terminated
Satisfaction Score on a Scale of 1 to 5
5
4
3
2
1
Status


Employee Satisfaction vs. Engagement Survey by Employment Status
Employee Satisfaction
Engagement Survey
1
2
3
4
5
Employment Status
Active
Terminated for Cause
Voluntarily Terminated


Employee Performance by Status with Satisfaction
Performance Rating (4 = Excellent, 1 = Needs Improvement)
Status
Active
Terminated for Cause
Voluntarily Terminated
Employee Satisfaction
5
4
3
2
1


count
10
5
0
0
10
20
30
40
ManagerID


count
20
15
10
5
0
0
10
20
30
40
ManagerID

The graphs indicate a relationship between employee performance and status, and employee satisfaction and status. There may be a small effect of salary on employement status. It seems that certain managers (ID = 11, 20, 39) managed a lot of employees who left, and there is a more even distribution of employees across managers. The chi-squared test below reveals that manager is a factor that can predict who will depart.

```
left_manager = table(hrdata$VoluntarilyTerminated, hrdata$ManagerName)

c2 = chisq.test(left_manager)

print(c2)
```

```
##
##  Pearson's Chi-squared test
##
## data:  left_manager
## X-squared = 46.301, df = 20, p-value = 0.0007328
```

The following correlation table will help identify variables that influence employee departure.

```
target_variable = hrdata$VoluntarilyTerminated
other_variables = hrdata %>% select(-VoluntarilyTerminated, -EmpStatusID, -Active, -TerminatedFo
rCause, -LastPerformanceReview_Date, -DOB, -TermReason, -Termd, -Employee_Name, -EmpID, -DateofT
ermination) %>% select_if(is.numeric)

correlations = sapply(other_variables, function(x) cor(target_variable, x, use = "complete.ob
s"))

correlation_df = data.frame(Variable = names(correlations), Correlation = correlations)
print(correlation_df)
```

```
##                                   Variable  Correlation
## MarriedID                         MarriedID  0.071630493
## MaritalStatusID             MaritalStatusID  0.118968544
## GenderID                           GenderID -0.017273957
## DeptID                               DeptID  0.100556025
## PerfScoreID                     PerfScoreID -0.024593128
## FromDiversityJobFairID FromDiversityJobFairID  0.191344197
## Salary                               Salary -0.126844142
## PositionID                       PositionID  0.191383553
## Zip                                     Zip -0.134312695
## ManagerID                       ManagerID  0.248577386
## EngagementSurvey           EngagementSurvey  0.055937276
## EmpSatisfaction             EmpSatisfaction  0.004880041
## SpecialProjectsCount   SpecialProjectsCount -0.189418482
## DaysLateLast30               DaysLateLast30 -0.002770520
## Absences                           Absences  0.077040402
## InState                             InState  0.155937193
## ProductionTech               ProductionTech  0.237335918
```

There are not many variables with high correlations.The highest seem to be the employee's manager, being married, whether they are hired at a diversity job fair, whether they live in the same state as the company headquarters, the number of special products they were assigned, and whether they're a Production Technician. I will experiment with three logistic regressions to predict whether an employee will leave or stay.

Models

Logistic Regression Models

I built three models. The first one using only the factors identified in the correlation table. The second model attempts to purposely overfit, but then the variables that seem to have a statistically significant relationship are used for the thrid modle.

```
binomial1 = glm(VoluntarilyTerminated ~ ManagerName + MarriedID + SpecialProjectsCount + InState
+ FromDiversityJobFairID + ProductionTech, data = hrdata, family = binomial)

summary(binomial1)
```

```
##
## Call:
## glm(formula = VoluntarilyTerminated ~ ManagerName + MarriedID +
##     SpecialProjectsCount + InState + FromDiversityJobFairID +
##     ProductionTech, family = binomial, data = hrdata)
##
## Coefficients: (1 not defined because of singularities)
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     0.01082    2.37945   0.005 0.996372
## ManagerNameAmy Dunn            -0.66983    2.12296  -0.316 0.752369
## ManagerNameBoard of Directors  -18.49038 2717.55919  -0.007 0.994571
## ManagerNameBrandon R. LeBlanc   -1.39315    1.53581  -0.907 0.364349
## ManagerNameBrannon Miller       -2.48237    2.09704  -1.184 0.236514
## ManagerNameBrian Champaigne    -14.82409 1354.02713  -0.011 0.991265
## ManagerNameDavid Stanley        -2.13972    2.12895  -1.005 0.314870
## ManagerNameDebra Houlihan       -1.75246    2.57105  -0.682 0.495486
## ManagerNameElijiah Gray         -1.98771    2.12248  -0.937 0.349014
## ManagerNameEric Dougall        -15.27230 1856.00344  -0.008 0.993435
## ManagerNameJanet King           -2.36784    2.13880  -1.107 0.268254
## ManagerNameJennifer Zamora       0.20187    1.62333   0.124 0.901035
## ManagerNameJohn Smith           -2.18563    2.50482  -0.873 0.382899
## ManagerNameKelley Spirea        -2.93725    2.13414  -1.376 0.168723
## ManagerNameKetsia Liebig        -2.63854    2.14080  -1.232 0.217763
## ManagerNameKissy Sullivan       -1.40219    2.11214  -0.664 0.506772
## ManagerNameLynn Daneault       -17.87951 1078.78094  -0.017 0.986777
## ManagerNameMichael Albert       -1.62385    2.11682  -0.767 0.443011
## ManagerNamePeter Monroe          0.06650    1.55518   0.043 0.965893
## ManagerNameSimon Roup            1.33388    1.26961   1.051 0.293432
## ManagerNameWebster Butler       -0.86229    2.11607  -0.407 0.683644
## MarriedID                        0.51966    0.29932   1.736 0.082546 .
## SpecialProjectsCount            -0.65308    0.44132  -1.480 0.138920
## InState                          0.90523    1.35178   0.670 0.503074
## FromDiversityJobFairID           1.69224    0.49789   3.399 0.000677 ***
## ProductionTech                        NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 370.54  on 310  degrees of freedom
## Residual deviance: 299.35  on 286  degrees of freedom
## AIC: 349.35
##
## Number of Fisher Scoring iterations: 16
```

```
binomial2 = glm(VoluntarilyTerminated ~ ManagerName + Absences + GenderID +
      InState + MarriedID + FromDiversityJobFairID + EngagementSurvey + EmpSatisfaction +
        DaysLateLast30 + Absences + RaceDesc + HispanicLatino +
      Salary + SpecialProjectsCount + PerfScoreID, data = hrdata, family = binomial)

summary(binomial2)
```

```
##
## Call:
## glm(formula = VoluntarilyTerminated ~ ManagerName + Absences +
##     GenderID + InState + MarriedID + FromDiversityJobFairID +
##     EngagementSurvey + EmpSatisfaction + DaysLateLast30 + Absences +
##     RaceDesc + HispanicLatino + Salary + SpecialProjectsCount +
##     PerfScoreID, family = binomial, data = hrdata)
##
## Coefficients: (1 not defined because of singularities)
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -1.574e+01  5.994e+03  -0.003 0.997904
## ManagerNameAmy Dunn            -4.610e-01  2.263e+00  -0.204 0.838584
## ManagerNameBoard of Directors  -1.884e+01  3.956e+03  -0.005 0.996201
## ManagerNameBrandon R. LeBlanc  -1.540e+00  1.634e+00  -0.943 0.345799
## ManagerNameBrannon Miller      -2.418e+00  2.193e+00  -1.102 0.270284
## ManagerNameBrian Champaigne    -1.478e+01  1.372e+03  -0.011 0.991404
## ManagerNameDavid Stanley       -2.152e+00  2.268e+00  -0.949 0.342861
## ManagerNameDebra Houlihan      -1.813e+00  2.687e+00  -0.675 0.499848
## ManagerNameElijiah Gray        -1.837e+00  2.242e+00  -0.820 0.412415
## ManagerNameEric Dougall        -1.518e+01  1.929e+03  -0.008 0.993721
## ManagerNameJanet King          -2.246e+00  2.178e+00  -1.031 0.302515
## ManagerNameJennifer Zamora      4.301e-01  1.766e+00   0.244 0.807580
## ManagerNameJohn Smith          -1.852e+00  2.738e+00  -0.676 0.498799
## ManagerNameKelley Spirea       -2.959e+00  2.250e+00  -1.315 0.188454
## ManagerNameKetsia Liebig       -2.506e+00  2.284e+00  -1.097 0.272498
## ManagerNameKissy Sullivan      -1.179e+00  2.250e+00  -0.524 0.600330
## ManagerNameLynn Daneault       -1.741e+01  1.074e+03  -0.016 0.987064
## ManagerNameMichael Albert      -1.644e+00  2.254e+00  -0.729 0.465871
## ManagerNamePeter Monroe        -1.936e-01  1.615e+00  -0.120 0.904630
## ManagerNameSimon Roup           1.286e+00  1.280e+00   1.004 0.315230
## ManagerNameWebster Butler      -9.172e-01  2.244e+00  -0.409 0.682723
## Absences                        2.373e-02  2.648e-02   0.896 0.370106
## GenderID                       -6.311e-03  3.096e-01  -0.020 0.983736
## InState                         9.114e-01  1.385e+00   0.658 0.510462
## MarriedID                       4.607e-01  3.127e-01   1.473 0.140751
## FromDiversityJobFairID          2.396e+00  6.736e-01   3.557 0.000376 ***
## EngagementSurvey                3.435e-01  2.398e-01   1.432 0.152129
## EmpSatisfaction                 4.981e-02  1.775e-01   0.281 0.778967
## DaysLateLast30                 -1.032e-01  1.823e-01  -0.566 0.571495
## RaceDescAsian                   1.653e+01  2.150e+03   0.008 0.993863
## RaceDescBlack or African American  1.580e+01  2.150e+03   0.007 0.994137
## RaceDescHispanic               -5.931e-01  7.182e+03   0.000 0.999934
## RaceDescTwo or more races       1.582e+01  2.150e+03   0.007 0.994128
## RaceDescWhite                   1.654e+01  2.150e+03   0.008 0.993860
## HispanicLatinoNo               -9.494e-01  5.595e+03   0.000 0.999865
## HispanicLatinoyes                      NA         NA      NA       NA
## HispanicLatinoYes              -1.262e+00  5.595e+03   0.000 0.999820
## Salary                         -5.854e-07  1.367e-05  -0.043 0.965830
## SpecialProjectsCount           -6.210e-01  4.351e-01  -1.427 0.153522
## PerfScoreID                    -5.128e-01  4.170e-01  -1.230 0.218808
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 370.54  on 310  degrees of freedom
## Residual deviance: 289.31  on 272  degrees of freedom
## AIC: 367.31
##
## Number of Fisher Scoring iterations: 16
```

```
binomial3 = glm(VoluntarilyTerminated ~ MarriedID + FromDiversityJobFairID +  ProductionTech, da
ta = hrdata, family = binomial)

summary(binomial3)
```

```
##
## Call:
## glm(formula = VoluntarilyTerminated ~ MarriedID + FromDiversityJobFairID +
##     ProductionTech, family = binomial, data = hrdata)
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -2.2321     0.3196  -6.984 2.87e-12 ***
## MarriedID                 0.4344     0.2703   1.607  0.10800
## FromDiversityJobFairID    1.5340     0.4282   3.583  0.00034 ***
## ProductionTech            1.3786     0.3182   4.333 1.47e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 370.54  on 310  degrees of freedom
## Residual deviance: 336.53  on 307  degrees of freedom
## AIC: 344.53
##
## Number of Fisher Scoring iterations: 4
```

Evaluating Logistic Regression Models

For a logistic regression, AIC measures can capture a balance between accuracy and simplicity. The third model has an AIC of 344, a bit better than the first model's AIC of 349, not not significant.

Using the models on test and training data and drawing ROC curves can help assess their performance as well.

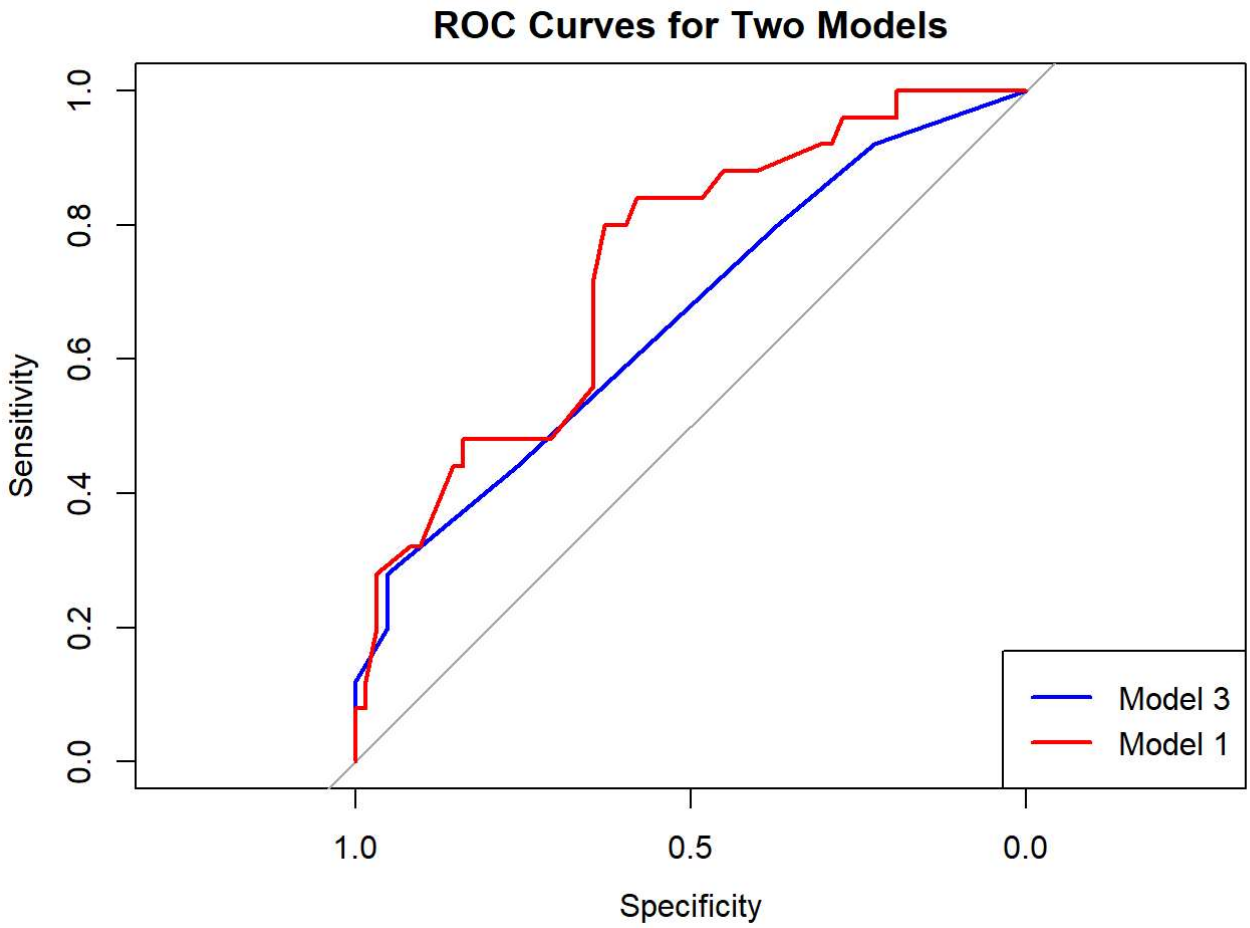
```
library(caret)
library(pROC)

#Establishing train and test data sets
set.seed(123)
ind = sample(2, nrow(hrdata), replace=TRUE, prob=c(0.7, 0.3))
train = hrdata[ind==1,]
test  = hrdata[ind==2,]

test$predicted_prob1 = predict(binomial1, newdata = test, type = "response")
test$predicted_class1 = ifelse(test$predicted_prob1 > 0.5, 1, 0)

confusion_matrix1 = table(test$VoluntarilyTerminated, test$predicted_class1)
print(confusion_matrix1)
```

```
##
##      0  1
##   0 52 10
##   1 14 11
```

```
test$predicted_prob3 = predict(binomial3, newdata = test, type = "response")
test$predicted_class3 = ifelse(test$predicted_prob1 > 0.5, 1, 0)

confusion_matrix3 = table(test$VoluntarilyTerminated, test$predicted_class3)
print(confusion_matrix3)
```

```
##
##      0  1
##   0 52 10
##   1 14 11
```

```
roc3 = roc(test$VoluntarilyTerminated, test$predicted_prob3)


roc1 = roc(test$VoluntarilyTerminated, test$predicted_prob1)


plot(roc3, col = "blue", main = "ROC Curves for Two Models")
lines(roc1, col = "red")
legend("bottomright", legend = c("Model 3", "Model 1"), col = c("blue", "red"), lwd = 2)
```



The models have identical false positive (10) and false negative (14) results. This indicates that removing managers and special projects from the model has little impact.

Yet the ROC curves indicate that the first model, with manager and special projects included, has improved sensitivity.

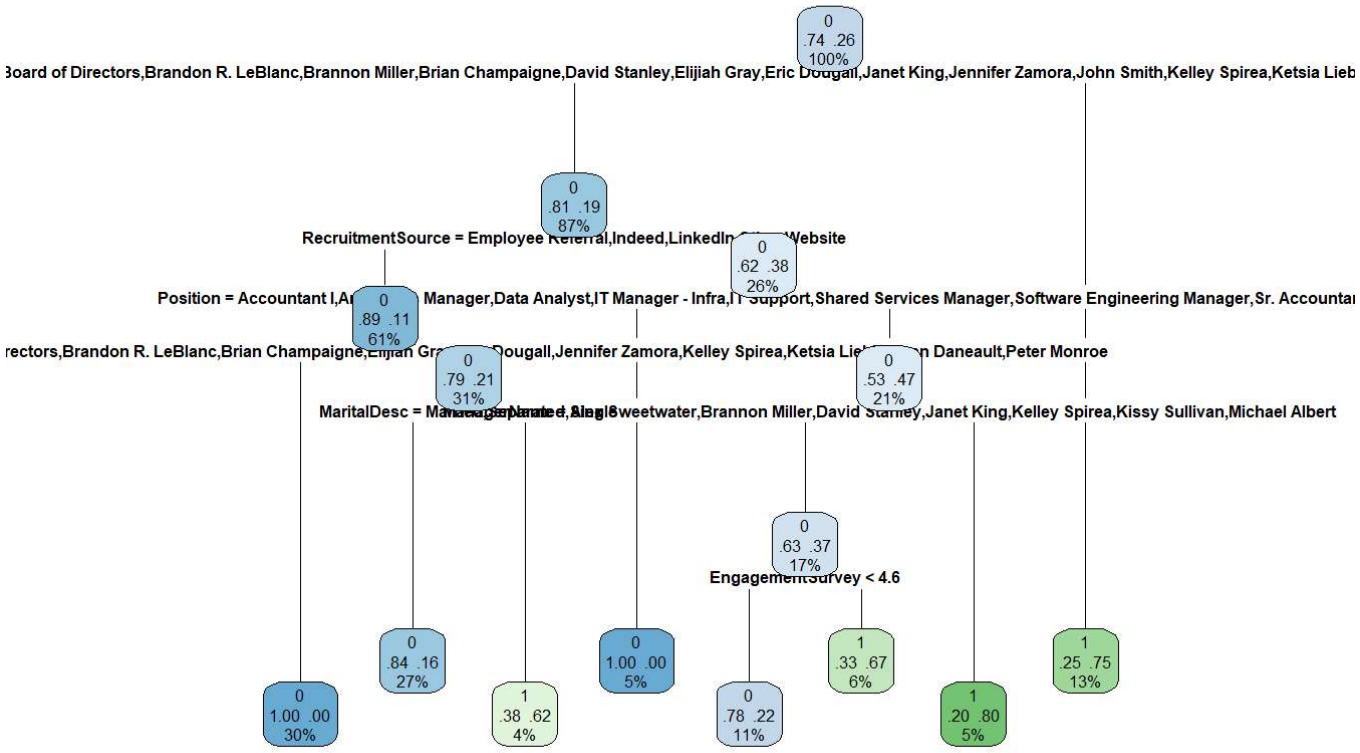
Unsupervised Learning with Decision Trees and Random Forests

Before attempting to create a decision tree model, I streamlined my data set by removing variables with dates and variables that were collinear with my target variable. Then I split my data into test and training sets, created the tree, and analyzed its performance.

```
library(mlr)
library(dplyr)
library(rpart)
library(rpart.plot)


#Getting rid of date columns, idiosyncratic values like names, and Employee Status variables
hrdata_nodate = hrdata %>% dplyr::select(
  -DOB, -DateofHire, -DateofTermination, -TermReason, -LastPerformanceReview_Date, -Termd, -Empl
oyee_Name, -EmploymentStatus, -Active, -TerminatedForCause, -EmpStatusID)

hrdata_nodate = na.omit(hrdata_nodate)
hrdata_nodate$VoluntarilyTerminated = as.factor(hrdata_nodate$VoluntarilyTerminated)

# Split data into training and testing sets
set.seed(123)
train_index = sample(1:nrow(hrdata_nodate), 0.7 * nrow(hrdata_nodate))
train_tree = hrdata_nodate[train_index, ]
test_tree = hrdata_nodate[-train_index, ]

tree_model = rpart(VoluntarilyTerminated ~ ., data=train_tree, method="class", control=rpart.con
trol(cp=0.01))
print(tree_model)
```

```
## n= 212
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 212 56 0 (0.7358491 0.2641509)
##    2) ManagerName=Alex Sweetwater,Board of Directors,Brandon R. LeBlanc,Brannon Miller,Brian Champaigne,David Stanley,Elijiah Gray,Eric Dougall,Janet King,Jennifer Zamora,John Smith,Kelley Spirea,Ketsia Liebig,Kissy Sullivan,Lynn Daneault,Michael Albert,Peter Monroe,Simon Roup 184 35 0 (0.8097826 0.1902174)
##      4) RecruitmentSource=Employee Referral,Indeed,LinkedIn,Other,Website 129 14 0 (0.8914729 0.1085271)
##        8) ManagerName=Alex Sweetwater,Board of Directors,Brandon R. LeBlanc,Brian Champaigne,Elijiah Gray,Eric Dougall,Jennifer Zamora,Kelley Spirea,Ketsia Liebig,Lynn Daneault,Peter Monroe 63  0 0 (1.0000000 0.0000000) *
##        9) ManagerName=Brannon Miller,David Stanley,Janet King,John Smith,Kissy Sullivan,Michael Albert,Simon Roup 66 14 0 (0.7878788 0.2121212)
##         18) MaritalDesc=Married,Separated,Single 58  9 0 (0.8448276 0.1551724) *
##         19) MaritalDesc=Divorced 8  3 1 (0.3750000 0.6250000) *
##      5) RecruitmentSource=CareerBuilder,Diversity Job Fair,Google Search 55 21 0 (0.6181818 0.3818182)
##       10) Position=Accountant I,Area Sales Manager,Data Analyst,IT Manager - Infra,IT Support,Shared Services Manager,Software Engineering Manager,Sr. Accountant 10  0 0 (1.0000000 0.0000000) *
##       11) Position=IT Manager - DB,Network Engineer,Production Manager,Production Technician I,Production Technician II,Software Engineer 45 21 0 (0.5333333 0.4666667)
##         22) ManagerName=Alex Sweetwater,Brannon Miller,David Stanley,Janet King,Kelley Spirea,Kissy Sullivan,Michael Albert 35 13 0 (0.6285714 0.3714286)
##           44) EngagementSurvey< 4.56 23  5 0 (0.7826087 0.2173913) *
##           45) EngagementSurvey>=4.56 12  4 1 (0.3333333 0.6666667) *
##         23) ManagerName=Elijiah Gray,Jennifer Zamora,Ketsia Liebig,Peter Monroe 10  2 1 (0.2000000 0.8000000) *
##    3) ManagerName=Amy Dunn,Debra Houlihan,Webster Butler 28  7 1 (0.2500000 0.7500000) *
```

```
rpart.plot(tree_model, type=2, extra=104, fallen.leaves=TRUE, cex=0.5, main="Decision Tree")
```




```
tree_predictions = predict(tree_model, test_tree, type="class")

conf_matrix = confusionMatrix(tree_predictions, test_tree$VoluntarilyTerminated)
print(conf_matrix)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 47 15
##          1 12 17
##
##                Accuracy : 0.7033
##                  95% CI : (0.5984, 0.7945)
##     No Information Rate : 0.6484
##     P-Value [Acc > NIR] : 0.1616
##
##                   Kappa : 0.335
##
##  Mcnemar's Test P-Value : 0.7003
##
##             Sensitivity : 0.7966
##             Specificity : 0.5312
##          Pos Pred Value : 0.7581
##          Neg Pred Value : 0.5862
##              Prevalence : 0.6484
##          Detection Rate : 0.5165
##    Detection Prevalence : 0.6813
##       Balanced Accuracy : 0.6639
##
##        'Positive' Class : 0
##
```

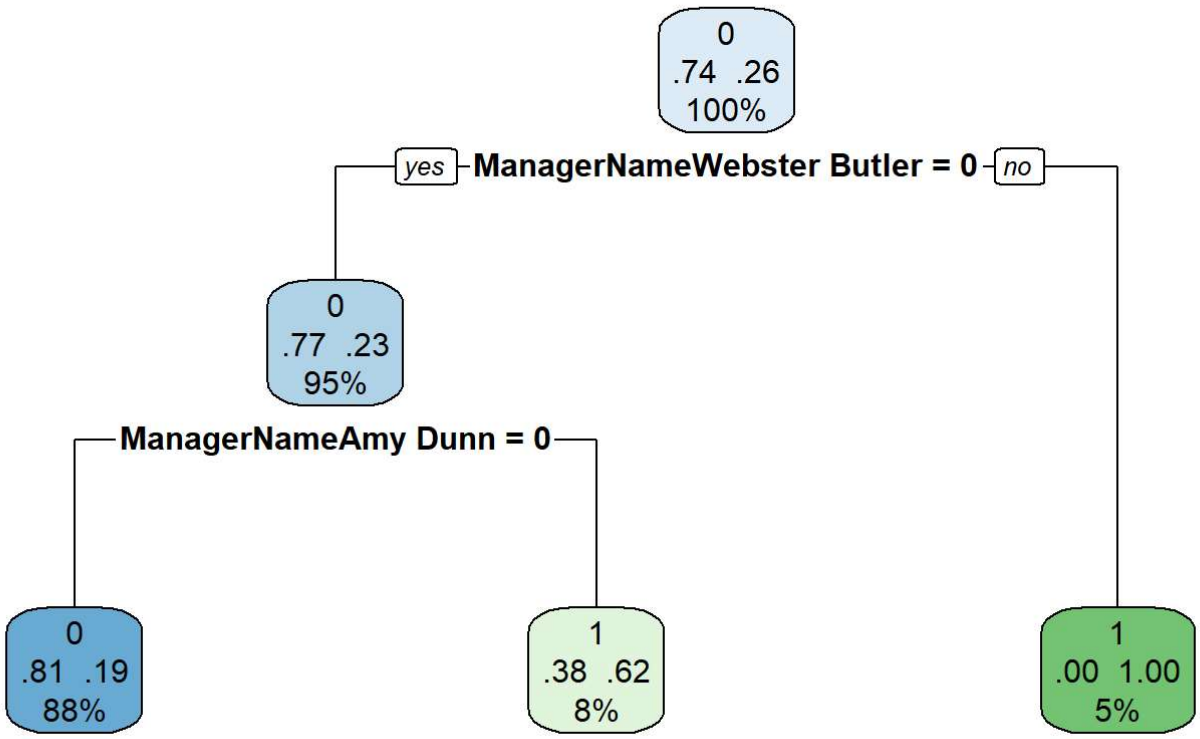
This decision tree had a slightly higher rate of false positives (17) and a slightly lower rate of false negatives (12) than my logistic models.

To try to improve the model, I used hyperparameter tuning.

```
train_control = trainControl(method="cv", number=10)
tree_cv_model = caret::train(VoluntarilyTerminated ~ ., data = train_tree, method = "rpart",
                             trControl = train_control)
print(tree_cv_model)
```

```
## CART
##
## 212 samples
##  29 predictor
##   2 classes: '0', '1'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 191, 190, 191, 190, 191, 191, ...
## Resampling results across tuning parameters:
##
##   cp          Accuracy   Kappa
##   0.01785714  0.7412771  0.25595840
##   0.07142857  0.7732900  0.23162843
##   0.17857143  0.7548701  0.08920188
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was cp = 0.07142857.
```

```
tune_grid = expand.grid(cp = seq(0.01, 0.1, by = 0.01))
tree_tuned_model = caret::train(VoluntarilyTerminated ~ ., data=train_tree, method="rpart", trCo
ntrol=train_control, tuneGrid=tune_grid)

rpart.plot(tree_tuned_model$finalModel, type=2, extra=104, fallen.leaves=TRUE, main="Tuned Decis
ion Tree")
```



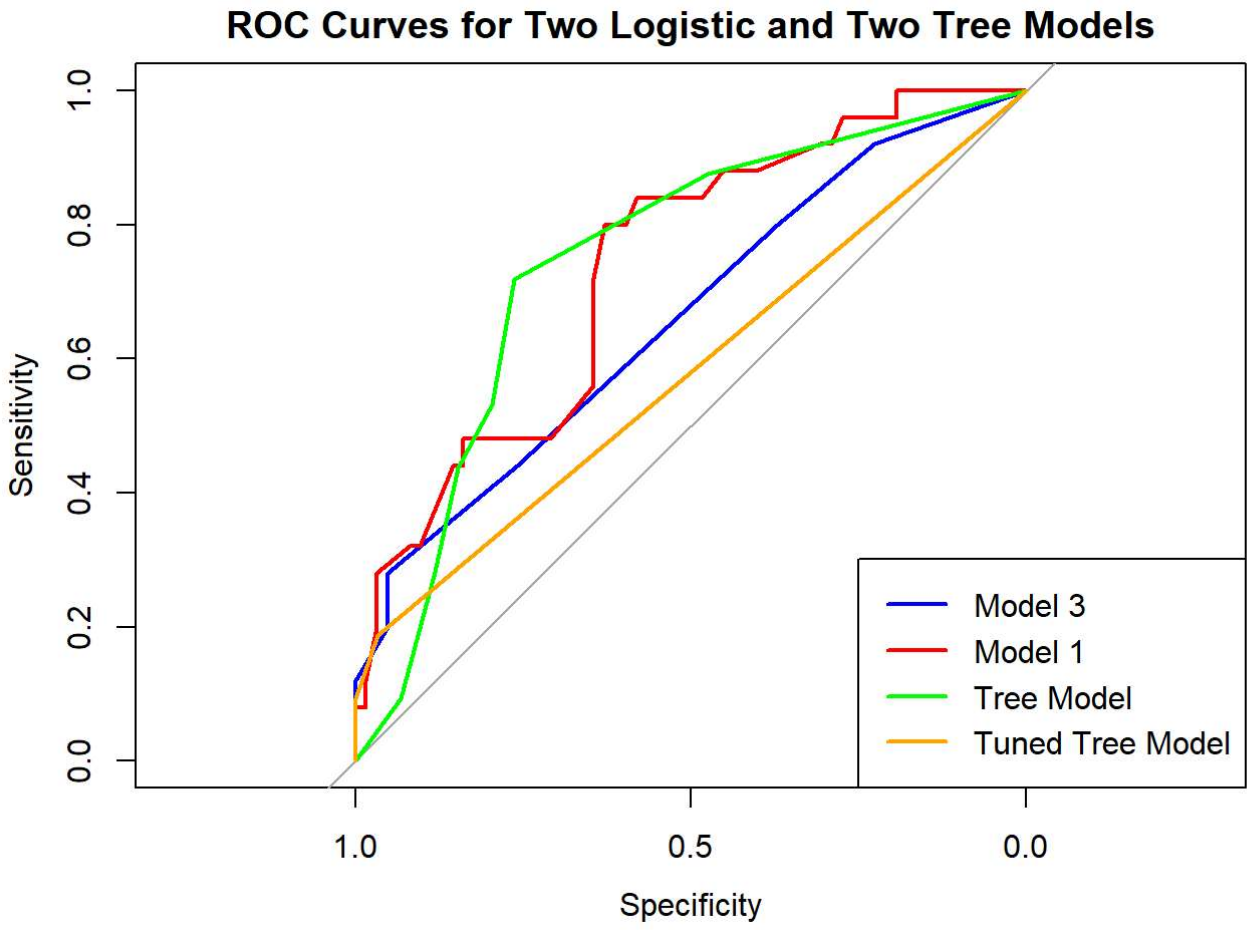

```
tree_predictions_tuned = predict(tree_tuned_model, test_tree, type="raw")

conf_matrix_tuned = confusionMatrix(tree_predictions_tuned, test_tree$VoluntarilyTerminated)
print(conf_matrix_tuned)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 57 26
##          1  2  6
##
##                Accuracy : 0.6923
##                  95% CI : (0.5868, 0.7849)
##     No Information Rate : 0.6484
##     P-Value [Acc > NIR] : 0.2225
##
##                   Kappa : 0.1854
##
##  Mcnemar's Test P-Value : 1.383e-05
##
##             Sensitivity : 0.9661
##             Specificity : 0.1875
##          Pos Pred Value : 0.6867
##          Neg Pred Value : 0.7500
##              Prevalence : 0.6484
##          Detection Rate : 0.6264
##    Detection Prevalence : 0.9121
##       Balanced Accuracy : 0.5768
##
##        'Positive' Class : 0
##
```

The tuned model only had two decision nodes, both determined by whether an employee had a certain manager. The tuned model dramatically decreased the number of false postitives (2) but increased the number of false negatives (26).

We can plot the ROC curves of both tree models and compare them to the logistic models. The AUC values and the ROC curves show that the first logistic model and the original tree model outperform the other models and perform quite similarly.

```
tree_prob_predictions = predict(tree_model, test_tree, type = "prob")
prob_positive_class = tree_prob_predictions[, 2]
roc_tree = roc(test_tree$VoluntarilyTerminated, prob_positive_class, levels = c("0", "1"))


tree_predictions_tuned = predict(tree_tuned_model, test_tree, type="prob")
prob_positive_tuned = tree_predictions_tuned[, 2]
roc_tree_tuned = roc(test_tree$VoluntarilyTerminated, prob_positive_tuned, levels = c("0", "1"))


plot(roc3, col = "blue", main = "ROC Curves for Two Logistic and Two Tree Models")
lines(roc1, col = "red")
lines(roc_tree, col = "green")
lines(roc_tree_tuned, col = "orange")
legend("bottomright", legend = c("Model 3", "Model 1", "Tree Model", "Tuned Tree Model"), col =
c("blue", "red", "green", "orange"), lwd = 2)
```




```
#AUC vales for all 4 models

AUCs = c(auc(roc1), auc(roc3), auc(roc_tree), auc(roc_tree_tuned))
auclabs = c("Logistic Model 1", "Logistic Model 3", "Tree Model", "Tuned Tree Model")

AUCvalues = data.frame(auclabs, AUCs)
print(AUCvalues)
```

```
##            auclabs      AUCs
## 1 Logistic Model 1 0.7354839
## 2 Logistic Model 3 0.6590323
## 3       Tree Model 0.7452331
## 4 Tuned Tree Model 0.5783898
```

Random Forest Model

Lastly, I attempted to construct a random forest model, and added its ROC curve and AUC measures to my comparisons.

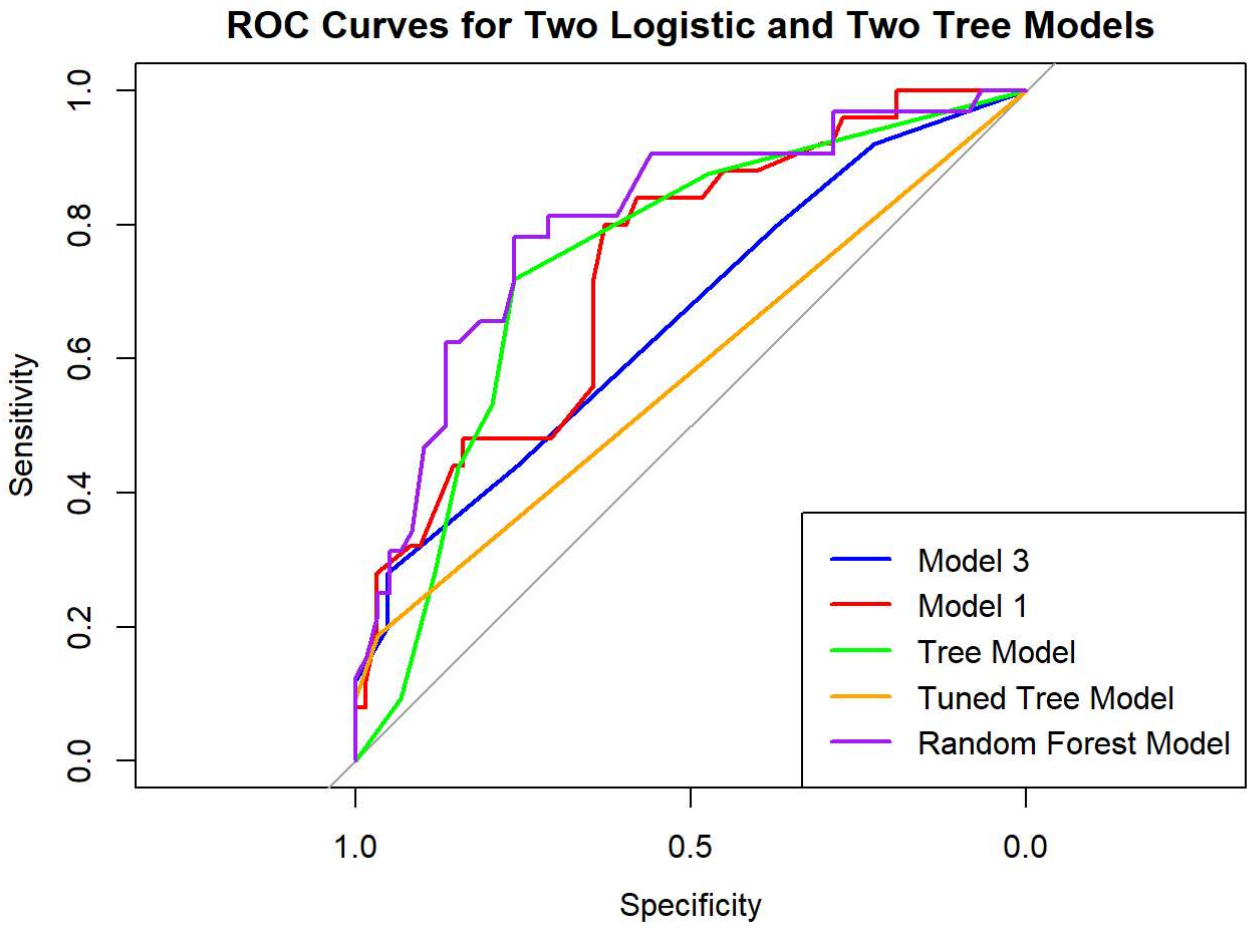
```
library(randomForest)
set.seed(71)

rf = randomForest(VoluntarilyTerminated~.,data=train_tree, ntree=100)
print(rf)
```

```
##
## Call:
##  randomForest(formula = VoluntarilyTerminated ~ ., data = train_tree,      ntree = 100)
##                Type of random forest: classification
##                      Number of trees: 100
## No. of variables tried at each split: 5
##
##         OOB estimate of  error rate: 26.89%
## Confusion matrix:
##     0  1 class.error
## 0 139 17   0.1089744
## 1  40 16   0.7142857
```

```
rf_predict = predict(rf, newdata = test_tree)

confusion_matrix_rf = confusionMatrix(test_tree$VoluntarilyTerminated, rf_predict)
print(confusion_matrix_rf)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 57  2
##          1 25  7
##
##                Accuracy : 0.7033
##                  95% CI : (0.5984, 0.7945)
##     No Information Rate : 0.9011
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.2212
##
##  Mcnemar's Test P-Value : 2.297e-05
##
##             Sensitivity : 0.6951
##             Specificity : 0.7778
##          Pos Pred Value : 0.9661
##          Neg Pred Value : 0.2187
##              Prevalence : 0.9011
##          Detection Rate : 0.6264
##    Detection Prevalence : 0.6484
##       Balanced Accuracy : 0.7364
##
##        'Positive' Class : 0
##
```

```
rf_predict_prob = predict(rf, newdata = test_tree, type = "prob")

prob_positive_class_rf = rf_predict_prob[, 2]


roc_rf <- roc(test_tree$VoluntarilyTerminated, prob_positive_class_rf, levels = c("0", "1"))

plot(roc3, col = "blue", main = "ROC Curves for Two Logistic and Two Tree Models")
lines(roc1, col = "red")
lines(roc_tree, col = "green")
lines(roc_tree_tuned, col = "orange")
lines(roc_rf, col = "purple")
legend("bottomright", legend = c("Model 3", "Model 1", "Tree Model", "Tuned Tree Model", "Random
Forest Model"), col = c("blue", "red", "green", "orange", "purple"), lwd = 2)
```




```
AUCs2 = c(auc(roc1), auc(roc3), auc(roc_tree), auc(roc_tree_tuned), auc(roc_rf))
auclabs2 = c("Logistic Model 1", "Logistic Model 3", "Tree Model", "Tuned Tree Model", "Random F
orest Model")

AUCvalues2 = data.frame(auclabs2, AUCs2)
print(AUCvalues2)
```

```
##              auclabs2     AUCs2
## 1    Logistic Model 1 0.7354839
## 2    Logistic Model 3 0.6590323
## 3          Tree Model 0.7452331
## 4    Tuned Tree Model 0.5783898
## 5 Random Forest Model 0.8034958
```

Here, the random forest model improves just slightly on the first logistic regression model. It may be useful to inform more explainable models, like a new logisitic model.

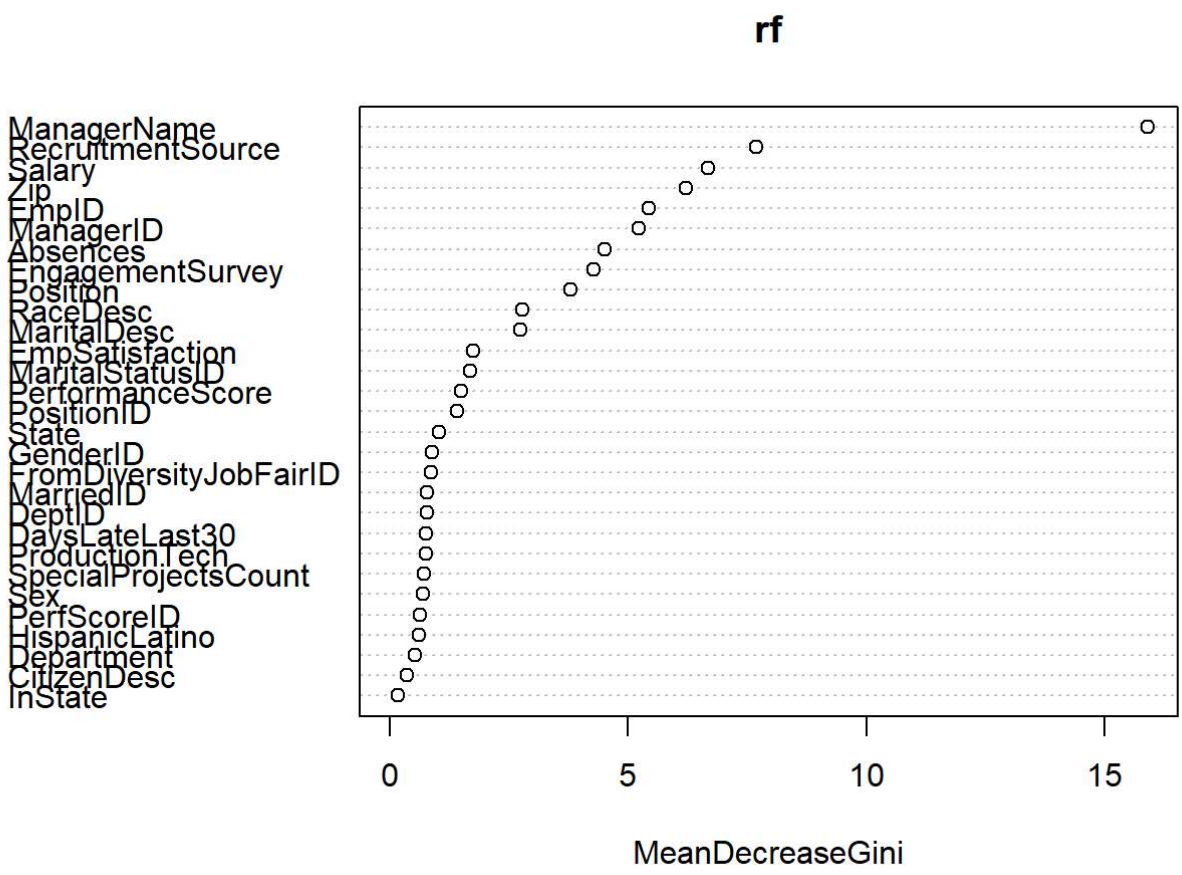
Feature Engineering with Random Forest

Finally, I used the random forest model to identify the 5 most important features and then created a logistic regression using them.

```
importance(rf)
```

```
##                         MeanDecreaseGini
## EmpID                          5.4348712
## MarriedID                      0.7756949
## MaritalStatusID                1.6880242
## GenderID                       0.8818978
## DeptID                         0.7683766
## PerfScoreID                    0.6230895
## FromDiversityJobFairID         0.8601842
## Salary                         6.6677563
## PositionID                     1.4101148
## Position                       3.7875088
## State                          1.0174688
## Zip                            6.2170172
## Sex                            0.6989870
## MaritalDesc                    2.7289382
## CitizenDesc                    0.3488450
## HispanicLatino                 0.6118941
## RaceDesc                       2.7692899
## Department                     0.5102558
## ManagerName                   15.9051456
## ManagerID                      5.2127057
## RecruitmentSource              7.6886957
## PerformanceScore               1.4917895
## EngagementSurvey               4.2628909
## EmpSatisfaction                1.7485203
## SpecialProjectsCount           0.7035532
## DaysLateLast30                 0.7624382
## Absences                       4.5043861
## InState                        0.1578542
## ProductionTech                 0.7503852
```

```
varImpPlot(rf)
```




```
binomial4 = glm(VoluntarilyTerminated ~ ManagerName + RecruitmentSource + Salary + Zip + Absence
s, data = hrdata, family = binomial)

summary(binomial4)
```

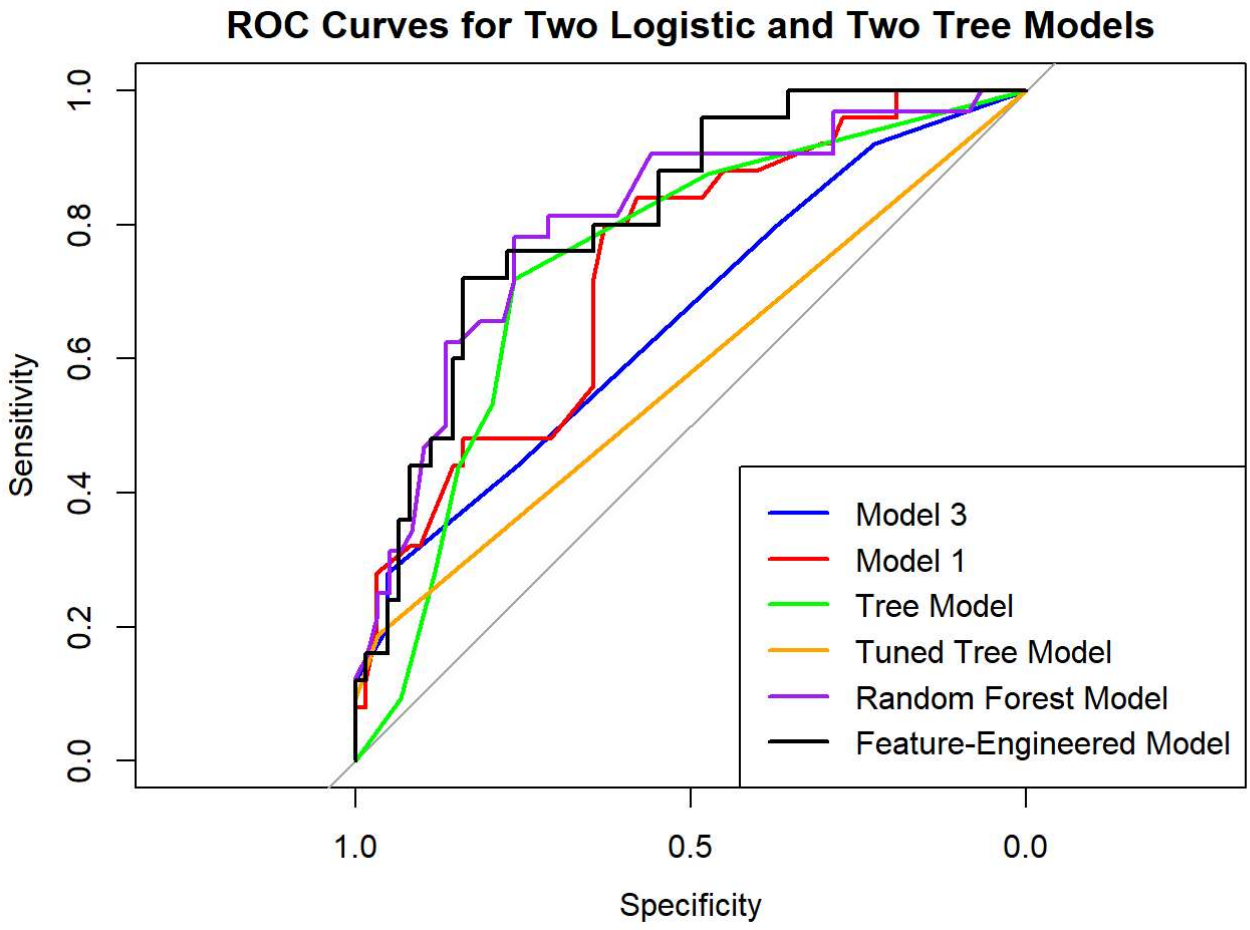
```
##
## Call:
## glm(formula = VoluntarilyTerminated ~ ManagerName + RecruitmentSource +
##     Salary + Zip + Absences, family = binomial, data = hrdata)
##
## Coefficients:
##                                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)                         -8.891e-01  1.555e+00  -0.572  0.56755
## ManagerNameAmy Dunn                  2.452e+00  1.141e+00   2.149  0.03162
## ManagerNameBoard of Directors       -3.296e+01  5.583e+03  -0.006  0.99529
## ManagerNameBrandon R. LeBlanc       -9.387e-01  1.520e+00  -0.618  0.53677
## ManagerNameBrannon Miller            4.219e-01  1.131e+00   0.373  0.70917
## ManagerNameBrian Champaigne         -1.659e+01  2.298e+03  -0.007  0.99424
## ManagerNameDavid Stanley             8.886e-01  1.165e+00   0.762  0.44579
## ManagerNameDebra Houlihan            3.414e-01  1.666e+00   0.205  0.83759
## ManagerNameElijiah Gray              7.403e-01  1.112e+00   0.666  0.50552
## ManagerNameEric Dougall             -1.726e+01  2.877e+03  -0.006  0.99521
## ManagerNameJanet King                8.086e-01  1.054e+00   0.767  0.44295
## ManagerNameJennifer Zamora          -5.364e-01  1.498e+00  -0.358  0.72026
## ManagerNameJohn Smith                4.104e-01  1.412e+00   0.291  0.77136
## ManagerNameKelley Spirea            -2.750e-01  1.172e+00  -0.235  0.81449
## ManagerNameKetsia Liebig             3.177e-01  1.188e+00   0.267  0.78915
## ManagerNameKissy Sullivan            1.308e+00  1.133e+00   1.154  0.24843
## ManagerNameLynn Daneault            -1.548e+01  1.662e+03  -0.009  0.99257
## ManagerNameMichael Albert            7.520e-01  1.136e+00   0.662  0.50792
## ManagerNamePeter Monroe             -3.903e-01  1.463e+00  -0.267  0.78959
## ManagerNameSimon Roup                1.359e+00  1.092e+00   1.245  0.21309
## ManagerNameWebster Butler            1.636e+00  1.146e+00   1.428  0.15324
## RecruitmentSourceDiversity Job Fair  8.053e-01  6.784e-01   1.187  0.23523
## RecruitmentSourceEmployee Referral  -2.671e+00  9.353e-01  -2.856  0.00429
## RecruitmentSourceGoogle Search      -3.224e-02  5.819e-01  -0.055  0.95581
## RecruitmentSourceIndeed             -1.134e+00  5.839e-01  -1.942  0.05212
## RecruitmentSourceLinkedIn           -1.486e+00  5.824e-01  -2.552  0.01072
## RecruitmentSourceOn-line Web application  1.892e+01  6.523e+03   0.003  0.99769
## RecruitmentSourceOther               1.659e+01  3.948e+03   0.004  0.99665
## RecruitmentSourceWebsite            -1.919e+00  1.262e+00  -1.521  0.12837
## Salary                              -3.589e-06  1.226e-05  -0.293  0.76980
## Zip                                 -2.372e-05  2.228e-05  -1.065  0.28691
## Absences                             3.520e-02  2.682e-02   1.312  0.18936
##
## (Intercept)
## ManagerNameAmy Dunn                 *
## ManagerNameBoard of Directors
## ManagerNameBrandon R. LeBlanc
## ManagerNameBrannon Miller
## ManagerNameBrian Champaigne
## ManagerNameDavid Stanley
## ManagerNameDebra Houlihan
## ManagerNameElijiah Gray
## ManagerNameEric Dougall
## ManagerNameJanet King
## ManagerNameJennifer Zamora
```

```
## ManagerNameJohn Smith
## ManagerNameKelley Spirea
## ManagerNameKetsia Liebig
## ManagerNameKissy Sullivan
## ManagerNameLynn Daneault
## ManagerNameMichael Albert
## ManagerNamePeter Monroe
## ManagerNameSimon Roup
## ManagerNameWebster Butler
## RecruitmentSourceDiversity Job Fair
## RecruitmentSourceEmployee Referral       **
## RecruitmentSourceGoogle Search
## RecruitmentSourceIndeed                  .
## RecruitmentSourceLinkedIn                *
## RecruitmentSourceOn-line Web application
## RecruitmentSourceOther
## RecruitmentSourceWebsite
## Salary
## Zip
## Absences
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 370.54  on 310  degrees of freedom
## Residual deviance: 276.55  on 279  degrees of freedom
## AIC: 340.55
##
## Number of Fisher Scoring iterations: 17
```

```
test$predicted_prob4 = predict(binomial4, newdata = test, type = "response")
test$predicted_class4 = ifelse(test$predicted_prob4 > 0.5, 1, 0)

confusion_matrix4 = table(test$VoluntarilyTerminated, test$predicted_class4)
print(confusion_matrix4)
```

```
##
##      0  1
##   0 55  7
##   1 14 11
```

```
roc4 = roc(test$VoluntarilyTerminated, test$predicted_prob4)

plot(roc3, col = "blue", main = "ROC Curves for Two Logistic and Two Tree Models")
lines(roc1, col = "red")
lines(roc_tree, col = "green")
lines(roc_tree_tuned, col = "orange")
lines(roc_rf, col = "purple")
lines(roc4, col = "black")
legend("bottomright", legend = c("Model 3", "Model 1", "Tree Model", "Tuned Tree Model", "Random
Forest Model", "Feature-Engineered Model"), col = c("blue", "red", "green", "orange", "purple",
"black"), lwd = 2)
```




```
AUCs3 = c(auc(roc1), auc(roc3), auc(roc_tree), auc(roc_tree_tuned), auc(roc_rf), auc(roc4))
auclabs3 = c("Logistic Model 1", "Logistic Model 3", "Tree Model", "Tuned Tree Model", "Radom Fo
rest Model", "Feature-Engineered Model")

AUCvalues3 = data.frame(auclabs3, AUCs3)
print(AUCvalues3)
```

```
##                   auclabs3     AUCs3
## 1         Logistic Model 1 0.7354839
## 2         Logistic Model 3 0.6590323
## 3               Tree Model 0.7452331
## 4         Tuned Tree Model 0.5783898
## 5       Radom Forest Model 0.8034958
## 6 Feature-Engineered Model 0.8135484
```

Conclusions, Limitations, and Recommendations for Decision-Making

Several iterations of modeling demonstrated how difficult it is to predict with high accuracy whether an employee will choose to remain with the company. While we can build a model with an accuracy approaching 3/4s, the casual linkages between factors and the decision to leave are far from clear. An employee recruited from a certain source might be more likely to leave because that source (i.e. a diversity recruiting fair) might identify them as members of group which is in turn poorly treated as an employee, or might be part of a group (i.e. LinkedIn) which have advanced self-promotion skills. Our models here cannot uncover the causal mechanisms behind human action.

However, the models can suggest emphases for management. A manager seems to be the largest determinant of employee retention. Identifying and training excellent managers should be a priority for the organization. Salary, of course, plays a key role, as does location (likely proximity to the company's headquarters). Interestingly, abecenses are a key predictor of employee turnover, apart from performance. It may be helpful to identify increases in employee absences, not for discplinary action but to create intervention systems to increase their likelyhood of staying.