HR Attrition Predictions

Nicole Bartholow/Chase Henderson

12/2/2018

Introduction

AlBaCo Analytics recommended the purchase of Sierra as the perfect complement to InBev's portfolio. This report provides an analysis of Sierra's associates to form a strategy for retention.

One of the reasons InBev in merging with Sierra is their outstanding human capital. Their research group has produced an impressive array of crowd-pleasing beers with strong brand loyalty, and their sales team has grown their distribution from a local California beer, to a national presence in the just 10 years, necessitating their second brewery in North Carolina.

As we move forward with our merger with Sierra, it is critical that we retain this talent that created the company that attracted us. Using the data provided to us regarding Sierra's associates, we'll start by recapping the distribution across departments and roles and then we'll dig into recent attrition trends. With that, we have analyzed factors that lead to attrition at Sierra, and created a model to predict folks at risk for attrition in a stable environment, so we can choose to take action to mitigate that risk, or, alternatively, plan for their departure. On a longer term basis, we can use those factors to devise strategies to reduce attrition by addressing the symptoms.

As we will see in this report, there is no smoking gun for predicting individual attrition. But there are some specific risk factors that have straighforward salves.

We were provided with a snapshot Sierra associates in two separate data files of 1470 Sierra associates total. While we have Attrition status for all associates, we used the population of 1170 employees to create a model. We then test that model to predict the Attrition status of the other 300. For initial evaluation, we removed columns that were meaningless to attrition analysis either because they were not environmental or because they had no variation in value across the dataset. These columns are ID, Employee Count, Employee Number, Over 18, Standard Hours, and Rand.

On the whole, the model reflected a 16.1% attrition rate, but it is not clear over what time period.

Analysis of the Current Associate Population and the Entire Population

First we will take a quick look at the snapshot population of Sierra associates. Then we will review the entire population for the complete overview of the population we are analyzing, including those that have left the company.

The snapshot of Sierra shows a company with 1233 active associates, divided over three departments of Research & Development, Sales and HR. These departments are further distributed across nine Roles, and within those roles, associates are assigned their level. Roles have mostly three levels, the exception being Sales Representative, which is only Level 1 and Level 2.

Figure 1 shows the counts of active associates across roles and levels.

```
## Using Freq as value column: use value.var to override.
```

JobRole	Level 1	Level 2	Level 3	Level 4	Level 5	Total
Sales Representative	44	6	0	0	0	50
Human Resources	23	13	4	0	0	40
Laboratory Technician	144	51	2	0	0	197
Research Scientist	189	55	1	0	0	245
Sales Executive	0	197	62	10	0	269
Manufacturing Director	0	85	40	10	0	135
Healthcare Representative	0	75	39	8	0	122
Research Director	0	0	28	26	24	78
Manager	0	0	10	47	40	97

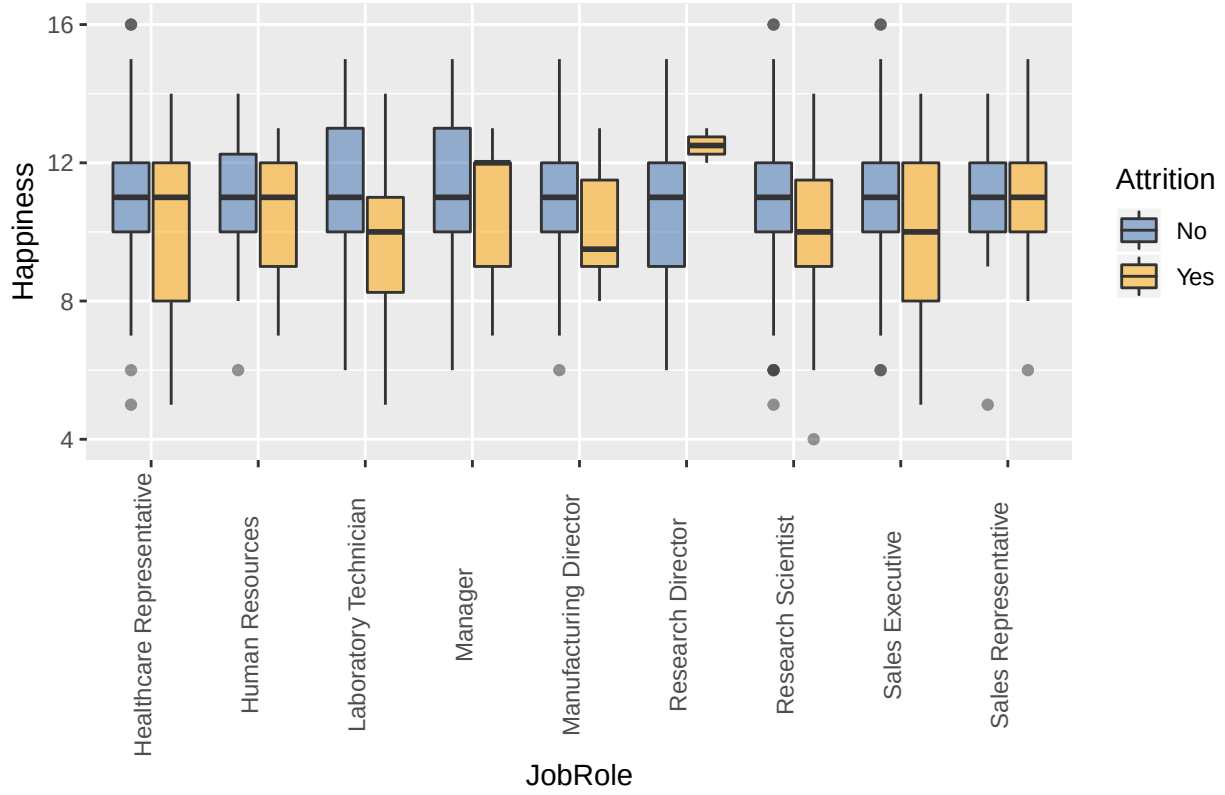
```
## Using Freq as value column: use value.var to override.
```

JobRole	Level 1	Level 2	Level 3	Level 4	Level 5	Total
Sales Representative	32	1	0	0	0	33
Human Resources	10	0	2	0	0	12
Laboratory Technician	56	5	1	0	0	62
Research Scientist	45	2	0	0	0	47
Sales Executive	0	36	17	4	0	57
Manufacturing Director	0	5	5	0	0	10
Healthcare Representative	0	3	5	1	0	9
Research Director	0	0	0	0	2	2
Manager	0	0	2	0	3	5

Figure 2 shows the counts of all associates across roles and levels in our dataset.

Sierra collects three measures of Satisfactions, Environment, Job and Relationship. It also collects a work/life balance rating. Each of these is a 1-4 rating. Surprisingly, our analysis shows these scores are not correlated, which is good because it means they aren't wasting time collecting double information. We added these scores together into one rating that we call Happiness. The idea behind Happiness is that each measure is independent, and if one of them is good, then it can make up for others that are below average. For example, if your Work/Life balance is poor, but your Environment Satisfaction is Very Good, then some balance is achieved. If the cumulative Happiness is low, then that is an indicator that an associate is missing a balancing positive to make up for negatives. The median and mean Happiness score for active associates - and for the total population - is right at 11 across all Job Roles. The median Happiness for those who have left is 10. Looking at the box plots in figure 3, we can see that Managers and Lab Technicians also post the median score of 11, but their upper quartiles are higher than other roles, at 13. The Happiness for the Attrition group is much more unpredictable, with the biggest populations, Lab Tech and Sales Executives and Research Scientist, posting lower ratings.

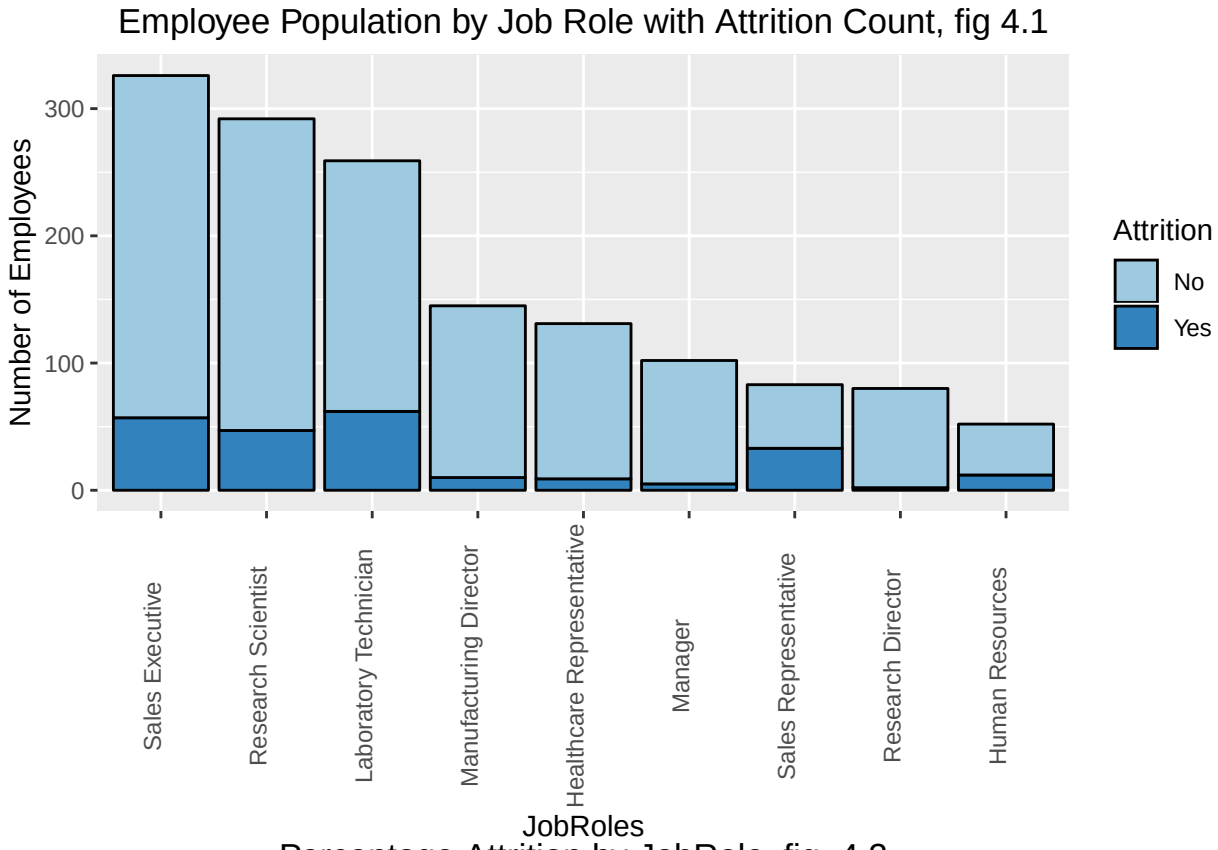
ımulative Happiness Score Across Departments − Current Associates, figure 3

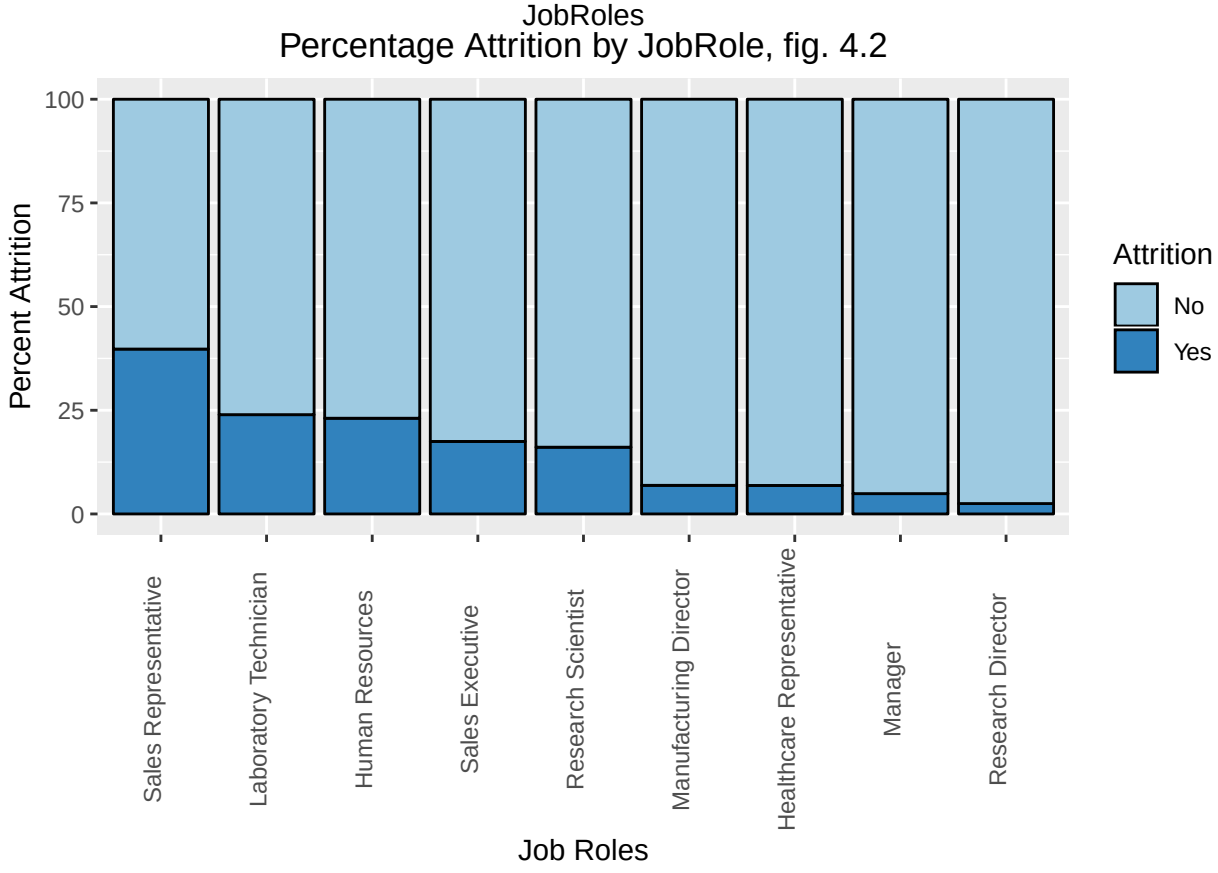


Turning our attention to the attrition data, it's important to look at the raw counts to assess what is costing the most money. But it's also important to look at percentages to formulate ideas about patterns and strategies, or drive deeper inquiry.

Figure 4.1 shows the relative population of each role and their attrition counts within that role. We can see that Laboratory Technicians and Sales Executives make up 50% of the attrition group with 62 and 57, respectively. Including Research Scientists and Sales Representatives covers 84% of the attrition group. Improving attrition in just these two groups provides an opportunity to significantly improve the turnover ratio overall.

Looking at the relative percentage in figure 4.2, we hope to identify if we should investigate more systemic issues within a group. Sales Representative attrition is significantly higher than Sales Executive, from a percentage standpoint, followed by Lab Technician, which we have already identified as a target for more investigation.

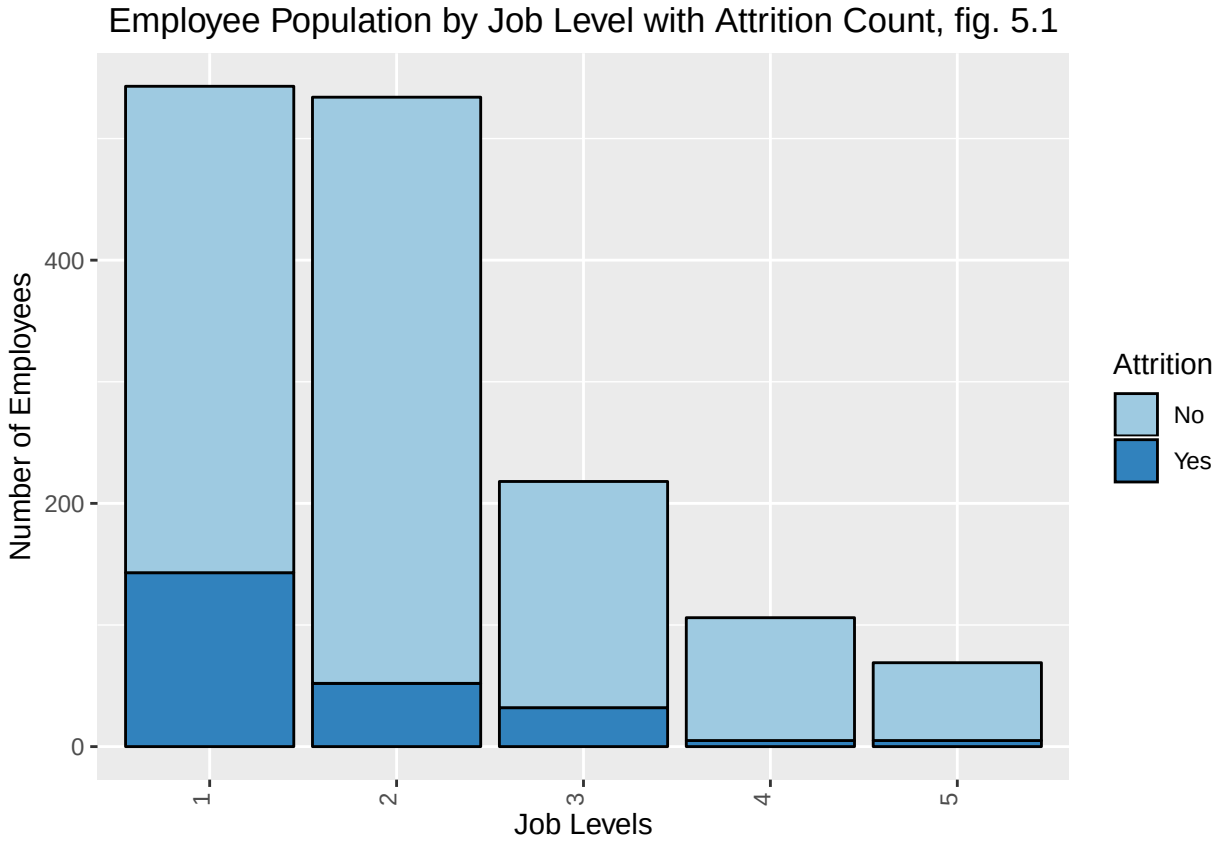

Employee Population by Job Role with Attrition Count, fig 4.1
Number of Employees
300
200
100
0
Sales Executive
Research Scientist
Laboratory Technician
Manufacturing Director
Healthcare Representative
Manager
Sales Representative
Research Director
Human Resources
JobRoles
Attrition
No
Yes

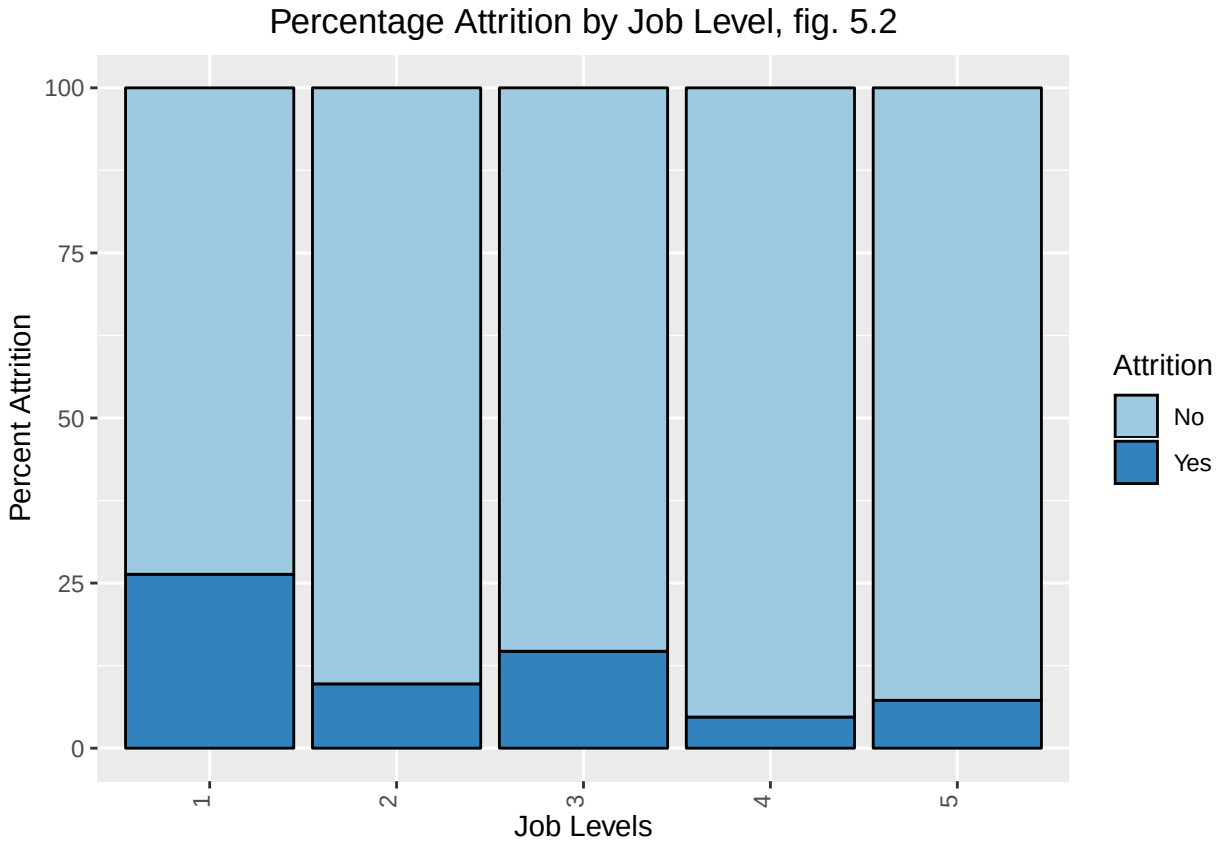

Percentage Attrition by JobRole, fig. 4.2
Percent Attrition
100
75
50
25
0
Sales Representative
Laboratory Technician
Human Resources
Sales Executive
Research Scientist
Manufacturing Director
Healthcare Representative
Manager
Research Director
Job Roles
Attrition
No
Yes

Changing our focus to look at job levels, Figure 5.1 shows the relative population of each level and their attrtion counts within that level. Note that the attrition is actually pushing Level 1 to look larger than Level 2, but in fact, in the stable population, Level 2 has 20% more associates. From the level perspective, it appears we would want to focus on Levels 1 and 2 from a pure turnover ratio perspective.

Looking at the relative percentage in figure 5.2, we see that Level 3 actually has a slightly higher turnover rate than Level 2.

Further investigation is warranted to identify how to prioritize Level 2 attrtion versus Level 3 attrition.





Predictors

We utilized four methodologies for identifying which features to use in our prediction models: logistic regression, graph & correlation analysis, and recursive feature elimination package functions. Our goal was for each to validate the findings of the other. We also used correlation plots to identify collinearity between features. Not surprisingly, all the Years and Age are correlated. Surprisingly, the Satisfactions do not show any correlation - that's why they should be added together and measured as a cumulative Happiness score. Also surprisingly, Happiness was not correlated to Attrition. Most surprisingly, these methodologies mostly did not concur on the best predictors.

We began by running a logistic regression to point us in the right direction. Logistic regression identified the following priorities: Overtime, Environment Satisfaction, Number of Companies Worked, Job Involvement, Business Travel, Years Since Last Promotion, Job Satisfaction, Distance from Home, Work Life Balance, and Marital Status. That list seems very logical and we embarked on validating it through our other methods.

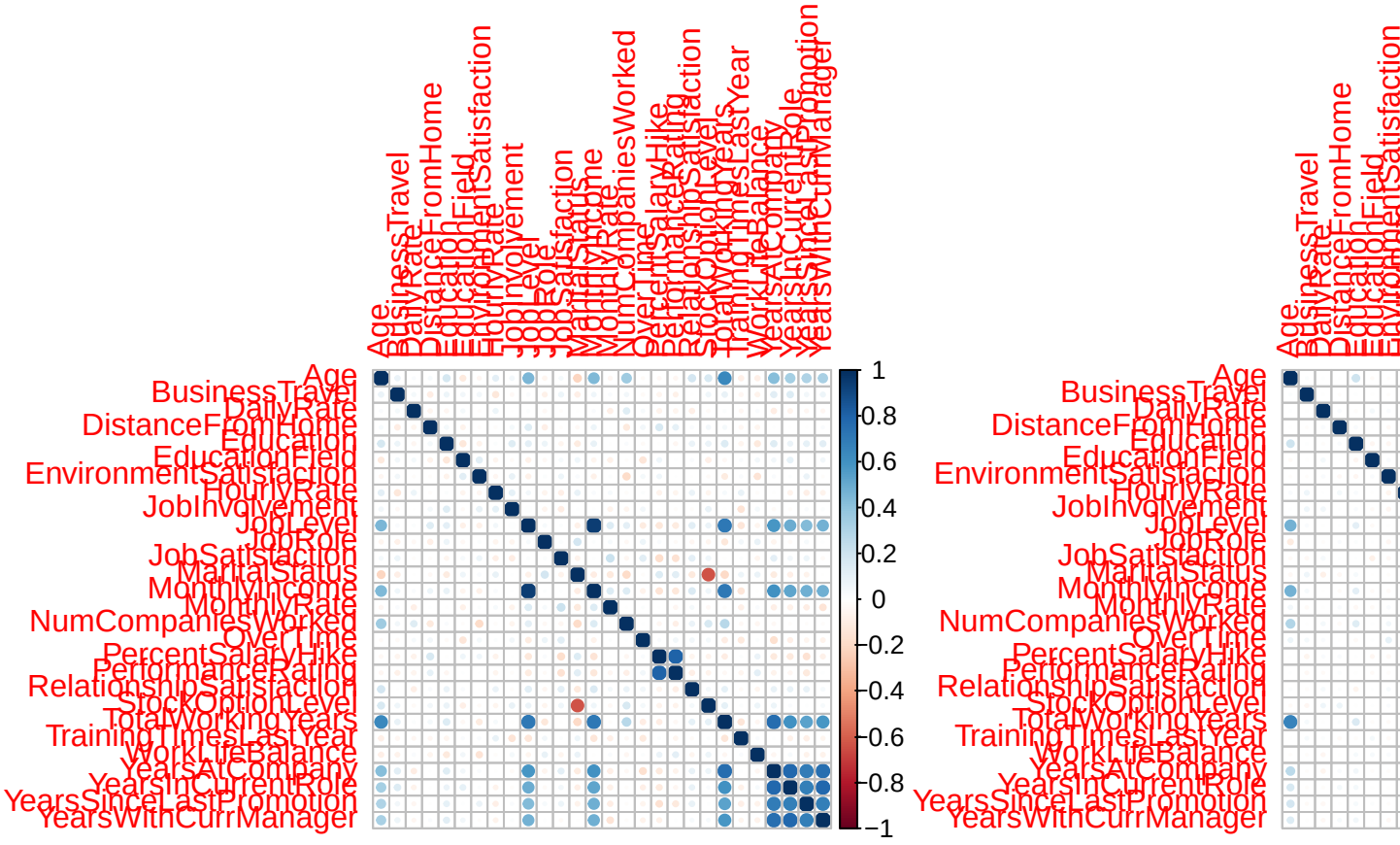
Direct Correlation did not often agree with the findings from logistic regression. Direct Correlation Analysis is more mundane than it sounds, running correlation for each variable against Attrtion to and comparing the value to the other results for ranking. Further analysis indicated that some of the items flagged by logistic regression were only specific values of a categorical variable. For example, Single versus not Single matters, but Married versus Divorced is not significant. So we created some derived binary variables out of continuous or categorigal variables to highlight the important transitions. This improved our correlation scores, but did not

improve predictions with the models that we used. In the end, Direct Correlation Analysis prioritized Overtime, Years at Company, Marital Status, Job Involvement and Environment Satisfaction.

Graphing showed Attrition relationships with Overtime, number of Companies Worked, Marital Status and Job Level. Other relationships were too subtle to pick up visually.

Because of the contradictions between the logistic regression, correlation and graphing, we decided to employ an automated feature selection function for validation. Initially we felt this muddies the water, buit directed us towards our best models by highlighting Job Level and all the measurements of Years, which had not shown strong correlation with Attrition.

Correlation Plots



Prediction Models

Neither of these models does a particularly good job of prediction. Each takes a different approach. kNN starts with the fewest variables and build up because each neighbor requirement can isolate the observation.

Naïve Bayes starts with a large number of variables and gets pruned down.

kNN Prediction Model

"Nearest Neighbors" models predicts based on "people like me". The drawback is that it gets overwhelmed with too many options and the number of "neighbors" quickly diminishes. Through several iterations guided by the different prediction selection methods, we maximized overall accuracy at .8616 useing: MaritalStatus, JobLevel, OverTime with k=7. This model predicted 9 correct Yes, with Sensitivity of .86 and Specificity of .82.

We felt a model which modestly errored on false positives was a better predictor. Our "Better Predictor of Yes"" uses: MaritalStatus, JobLevel, OverTime & Year1 with k= 7 nearest neighbors. With this model Overall Accuracy drops to 85% Accuracy, with 10 correct Yes predictions. This model had a Sensitivity : 0.8627 and Specificity : 0.6250.

```
## Confusion Matrix and Statistics
##
##
##       1   2
##   1 249   2
##   2  40   9
##
##                Accuracy : 0.86
##                  95% CI : (0.8155, 0.8972)
##     No Information Rate : 0.9633
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.2554
##  Mcnemar's Test P-Value : 1.135e-08
##
##             Sensitivity : 0.8616
##             Specificity : 0.8182
##          Pos Pred Value : 0.9920
##          Neg Pred Value : 0.1837
##              Prevalence : 0.9633
##          Detection Rate : 0.8300
##    Detection Prevalence : 0.8367
##       Balanced Accuracy : 0.8399
##
##        'Positive' Class : 1
##

## Confusion Matrix and Statistics
##
##    kNN_Prediction
##       1   2
##   1 249   2
##   2  40   9
##
##                Accuracy : 0.86
##                  95% CI : (0.8155, 0.8972)
##     No Information Rate : 0.9633
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.2554
##  Mcnemar's Test P-Value : 1.135e-08
```

```
##
##             Sensitivity : 0.8616
##             Specificity : 0.8182
##          Pos Pred Value : 0.9920
##          Neg Pred Value : 0.1837
##              Prevalence : 0.9633
##          Detection Rate : 0.8300
##    Detection Prevalence : 0.8367
##       Balanced Accuracy : 0.8399
##
##        'Positive' Class : 1
##
```

```
## Confusion Matrix and Statistics
##
##
##       1   2
##   1 249   2
##   2  40   9
##
##                Accuracy : 0.86
##                  95% CI : (0.8155, 0.8972)
##     No Information Rate : 0.9633
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.2554
##  Mcnemar's Test P-Value : 1.135e-08
##
##             Sensitivity : 0.8616
##             Specificity : 0.8182
##          Pos Pred Value : 0.9920
##          Neg Pred Value : 0.1837
##              Prevalence : 0.9633
##          Detection Rate : 0.8300
##    Detection Prevalence : 0.8367
##       Balanced Accuracy : 0.8399
##
##        'Positive' Class : 1
##
```

Naive Bayes Prediction Model

The naive Bayes model calculates probabilities for each predictor from a learning dataset and applies those probabilities to the test set to make predictions. With naive Bayes, it's simple to start with all the features as possible predictors and prune down the model based on accuracy.

The highest Accuracy Model uses OverTime, YearsAtCompany, YearsInCurrentRole, YearsWithCurrManager, JobLevel. This model achieved 87% Accuracy overall, with nineteen correct Yes, but thirty-one false positives. Also of note is the Sensitivity of .97 and Specificity of .37. In our case of predicting Attrition, the cost of a false positives, or Type I error, is lower than the cost of mis-diagnosing a yes as a no, or Type II error. For our purposes, a model predicting more Yes is generally better, unless it is presuming everyone is a Yes.

```
##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace)
##
## A-priori probabilities:
## Y
##        No       Yes
## 0.8393162 0.1606838
##
## Conditional probabilities:
##      OverTime
## Y            No       Yes
##   No  0.7606925 0.2393075
##   Yes 0.4680851 0.5319149
##
##      YearsAtCompany
## Y         [,1]     [,2]
##   No  7.343177 6.075246
##   Yes 5.335106 6.048772
##
##      YearsInCurrentRole
## Y         [,1]     [,2]
##   No  4.510183 3.681981
##   Yes 3.053191 3.225457
##
##      YearsWithCurrManager
## Y         [,1]     [,2]
##   No  4.376782 3.604645
##   Yes 3.026596 3.229929
##
##      JobLevel
## Y         [,1]      [,2]
##   No  2.124236 1.1128218
##   Yes 1.670213 0.9409098
```

```
## Confusion Matrix and Statistics
##
##
## New_NB_Prediction4  No Yes
##                No  910 132
##                Yes  72  56
##
##                Accuracy : 0.8256
##                  95% CI : (0.8027, 0.847)
##     No Information Rate : 0.8393
##     P-Value [Acc > NIR] : 0.9045
##
##                   Kappa : 0.2578
##  Mcnemar's Test P-Value : 3.615e-05
##
##             Sensitivity : 0.9267
##             Specificity : 0.2979
```

```
##          Pos Pred Value : 0.8733
##          Neg Pred Value : 0.4375
##              Prevalence : 0.8393
##          Detection Rate : 0.7778
##    Detection Prevalence : 0.8906
##       Balanced Accuracy : 0.6123
##
##        'Positive' Class : No
##
```

```
## Confusion Matrix and Statistics
##
##
## Test_Data_NB_Pred4  No Yes
##                No  243  31
##                Yes   8  18
##
##                Accuracy : 0.87
##                  95% CI : (0.8266, 0.9059)
##     No Information Rate : 0.8367
##     P-Value [Acc > NIR] : 0.065830
##
##                   Kappa : 0.4136
##  Mcnemar's Test P-Value : 0.000427
##
##             Sensitivity : 0.9681
##             Specificity : 0.3673
##          Pos Pred Value : 0.8869
##          Neg Pred Value : 0.6923
##              Prevalence : 0.8367
##          Detection Rate : 0.8100
##    Detection Prevalence : 0.9133
##       Balanced Accuracy : 0.6677
##
##        'Positive' Class : No
##
```

Recommendations Summary

Neither Model is particularly great at predicting attrition, but the analysis of the trends gives insights and threads to further study, and perhaps a broader array of models.

Further investigation is warranted. The high Level 1 turnover rate should be benchmarked against other jobs of their type, and also should be quantified in terms of hiring and training expenditure. Level 1 roles had a high percentage of Single associates with little work experience.

Analyzing the jobs and levels together, it's also important to view larger opportunity costs with turnover

Further investigation is warranted to identify how to prioritize Level 2 versus Level 3 attrition.

Other References

Youtube PowerPoint Presentation

https://youtu.be/X7wWLbDmeDM

Github Respository

https://github.com/NicoleABartholow/MSDS6306CaseStudy2