Executive Summary

Aaron Camacho

Problem Statement

"People are companies' most important assets" (Altman, 2017) and the cost to replace those assets is astonishing. "Josh Bersis of Deloitte believes the cost of losing an employee can range from tens of thousands of dollars to 1.5 – 2.0x the employee's annual salary" (Altman, 2017). If it is possible to predict which employees will leave the company and the potential reasons why, it can save companies hundreds of thousands of dollars. Also, by identifying the reasons why an employee might leave the company, the analysis can direct the company to make the necessary changes to increase employee morale and make a happier and more productive work place.

The question we want to answer is: What are the most likely reasons for an employee to leave a company and can the turnover possibility be predicted before it happens? Appearing simple at first, the question demands the consideration of various factors when attempting to determine whether one variable has a greater influence than another.

Hypothesis

If an employee has more tenure with the company, he/she is less likely to leave.

Data Analysis Summary

Logistic regression was used due to its ability to handle both categorical and numeric variables. However, other standard models were compared to discover the best approach (i.e. decision tree, random forest, CART, etc.).

The initial data set contained 35 variables in that could influence an employee to make a determination to leave the company. The data set incuded 1,470 rows each representing an employee. Out of the 1,470 employees 16.12 percent, or 237 individuals, left the company. Each variable was analyzed for how it related and impacted the variable Attrition. For instance, when the variable "YearsAtCompany" was analyzed, it was found that employees with more tenure were in fact less likely to leave (Figure 1). Though the comparison analysis of each variable gave insight and answered the hypothesis, it was believed that deeper understand could be achieved. New variables were created based off of conventions of the given variables. The following three variables were created:

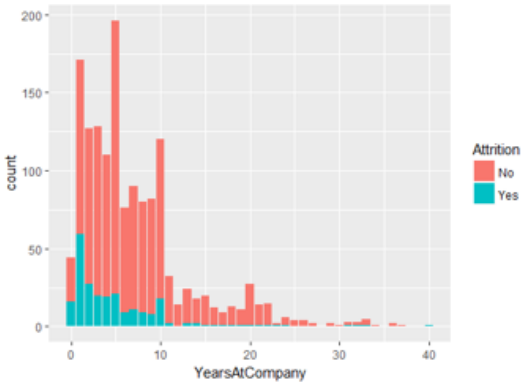


Figure 1: YearsAtCompany

Average Tenure per Job = TotalWorkingYears/NumCompaniesWroked. For this group the assumption was made that certain individuals are motivated by change. These individuals work for a company for a few years but end moving to a new company within a few years. We will see that those with a lower average tenure per a job are more apt to leave.

Years without Promotion in Current Role = YearsInCurrentRole – YearsSinceLastPromotion. The assumption here that was made here was employees that are seeking growth through a promotion are more likely to leave if a promotion is not gained within a reasonable amount of time. It is important to note that it is unclear whether the current role was a promotion or not.

Years without Promotion with Current Manager: YearsWithCurrentManager – YearsSinceLastPromotion. Like the variable above, the assumption made here was that employees that are seeking growth through a promotion are more likely to leave if a promotion is

not gained within a reasonable amount of time. However, it focuses on time with a manager vs current time in a role. It is important to note that it is possible that an employee might have a new manager but has not received a promotion which would cause a negative value.

Each variable was then checked for correlation, outlies, and missing values. Though there were no missing values found, there were many variables that were correlated and two variables that had a significant number of outliers, Monthly Income and Years At Company. Correlation between variables was found using the function **corplot()** in the R Programming Language (Figure 2). Highly correlated variables are variables that are approaching +/- 1. Variables that have a near zero variance are represented with a "?". Variables that are equal to. greater than, or less than 0.60/-0.60 respectively, were considered highly correlated and were removed from the data set. Also, variables that have a near zero variance and the two variables mentioned above with many outliers were also removed.

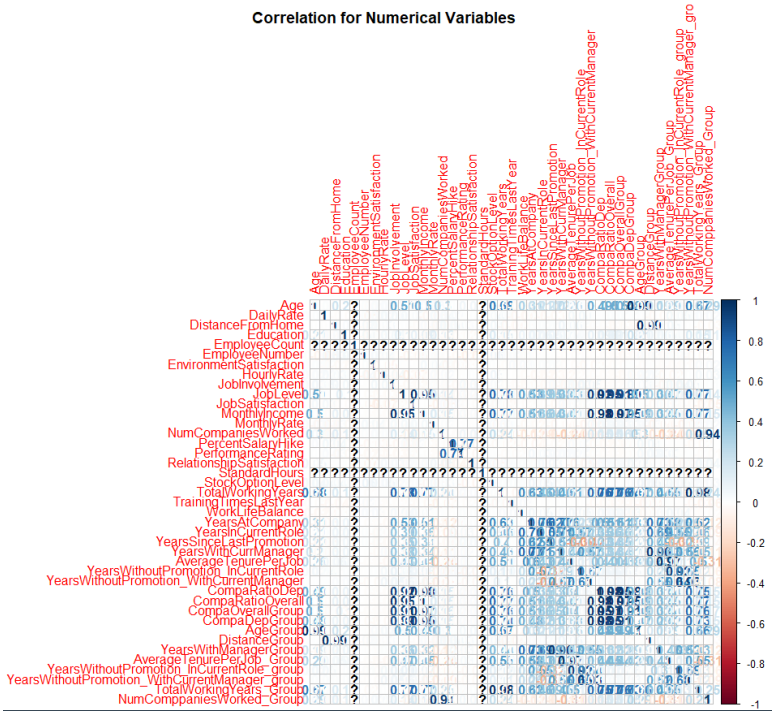


Fig. 2: Uncleaned data set with significant correlation

To assist in the analysis of the data set the following variables were binned/grouped to reduce the number of sub-categories and to gain a better understanding of how they affect attrition. Take for example the variable "Age". There are 43 ages contained within the variable. Though some understanding could be gained with so many ages, it made the analysis vague and difficult to interpret. Therefore, binning was performed on the variable "Age" (see diagram below). Binning the ages into groups allowed us to visualize certain age groups vs attrition level.

Age Group:

1 = 25 or less years

2 = 26 to 30

3 = 31 to 35

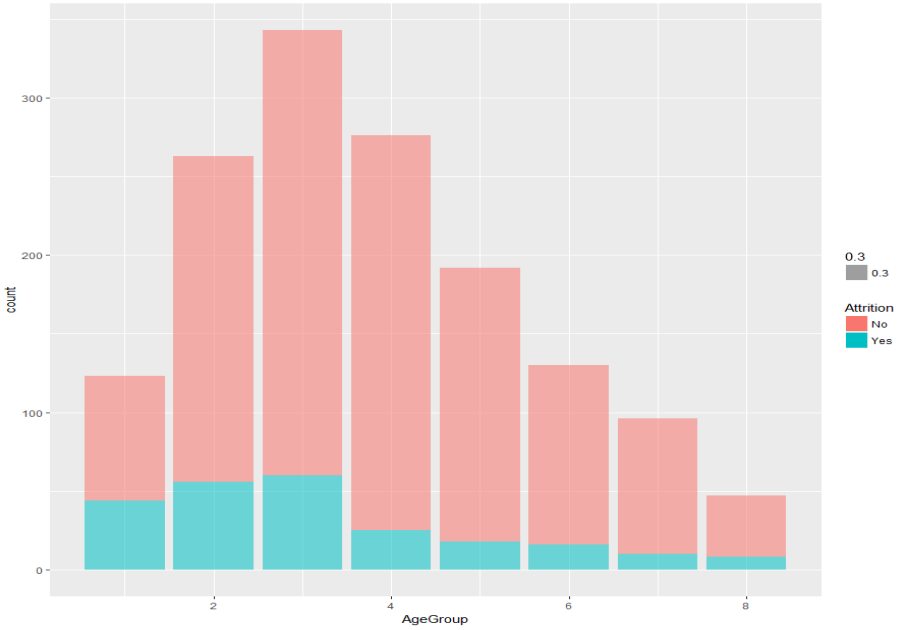
4 = 36 to 40

5 = 41 to 45

6 = 46 to 50

7 = 51 to 55

8 = 56 or greater



Beyond the variable Age, Distance from home, years with Manager, Average Tenure per a job, Years without promotion in current role and with current manager, Total working years, Number of companies worked for and years at company were all binned. All original variables that were binned were removed in favor of the new binned variables. However, NumCompaniesWorked was be kept instead of NumCompaniesWorked_Group because the p-value was lower in the original variable. After performing data analysis and data cleaning the following 22 variables were kept:

```
 [1] "Attrition"                          "BusinessTravel"
 [3] "Department"                         "Education"
 [5] "EducationField"                     "EnvironmentSatisfaction"
 [7] "Gender"                             "JobInvolvement"
 [9] "JobLevel"                           "JobRole"
[11] "JobSatisfaction"                    "MaritalStatus"
[13] "NumCompaniesWorked"                 "OverTime"
[15] "PercentSalaryHike"                  "RelationshipSatisfaction"
[17] "TrainingTimesLastYear"              "WorkLifeBalance"
[19] "DistanceGroup"                      "YearsWithManagerGroup"
[21] "AverageTenurePerJob_Group"          "YearsWithoutPromotion_WithCurrentManager_group"
```

The following methods were applied to test the data set and produce a predictive model: Comparative model testing, hierarchal clustering with k-modes, and an evaluative method logistic regression.

Techniques, Tools and Findings

Hierarchal Clustering K – Modes was chosen for the analytic method because of its ability to handle categorical data and represent which values occur most in the data set. PCA and other options lacked the ability to cope with the categorical data contained within the data. Three clusters were created out of the original data set to discover which units appear the most within each variable. The following was observed:

Business Travel: "Travel_Rarely" reoccurs the most in all three clusters.

Department: Research & Development reoccurs the most in all three clusters.

AverageTenturePerJob_Group: In two of three clusters group "2" or tenure of "2 to 3 years" is most reoccurring.

YearsWithManagerGroup: Group 1 or "2 or less years" reoccurs the most in all three clusters.

OverTime: "No" occurs most in all three clusters.

```
> cluster.results
K-modes clustering with 3 clusters of sizes 741, 431, 298

Cluster modes:
  Attrition BusinessTravel            Department Education EducationField EnvironmentSatisfaction Gender JobInvolvement JobLevel
1        No  Travel_Rarely Research & Development         3        Medical                      3   Male             3        2
2        No  Travel_Rarely Research & Development         3  Life Sciences                      3 Female             3        1
3        No  Travel_Rarely Research & Development         2  Life Sciences                      4   Male             2        2
            JobRole JobSatisfaction MaritalStatus NumCompaniesWorked OverTime PercentSalaryHike RelationshipSatisfaction
1    Sales Executive               3       Married                 1       No               12                        3
2 Research Scientist               4       Married                 1       No               13                        2
3    Sales Executive               4        Single                 1       No               14                        4
  TrainingTimesLastYear WorkLifeBalance DistanceGroup YearsWithManagerGroup AverageTenurePerJob_Group
1                     2               3             1                      1                        1
2                     3               3             1                      1                        2
3                     3               3             1                      1                        2
  YearsWithoutPromotion_WithCurrentManager_group
1                                              3
2                                              4
3                                              4
```

Comparative model testing was used to identify the analytical method that could potentially produce the most accurate results. With the cleaned and prepared data set, different analytic methods were compared against one another. What was found is that logistic regression gave the most accurate results when creating a predictive model (Figure 3). Therefore, logistic regression was chosen along with its ability handle both numerical and categorical data.

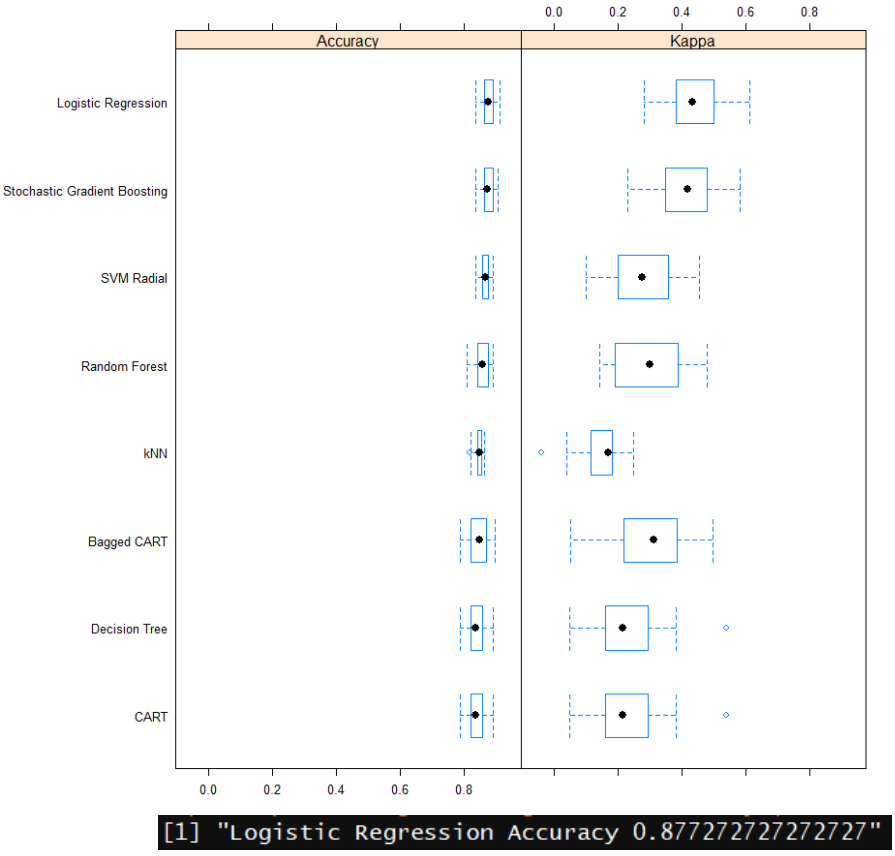


Figure 3: Comparative Model Testing

Logistic Regression:

We see in the logistic regression model and the Chi-Squared calculation that the critical value at 95 percent confidence and 992 degrees of freedom is 1,066.385. Since the residual deviance of 575.87 is less than the critical value the null model is not rejected. In other words, we have a reliable model at 95 percent confidence level. Also, we can see that the most significant variables are "Business Travel, Environmental Satisfaction, Job Involvement, Marital Status, Num of Companies Worked, Overtime, Years Without Promotion With Current Manager group" (Figure 4).

```
Coefficients:
                                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                                   -8.407e+00  1.615e+03  -0.005 0.995845
BusinessTravelTravel_Frequently                2.051e+00  5.424e-01   3.782 0.000156 ***
BusinessTravelTravel_Rarely                    1.195e+00  5.043e-01   2.369 0.017853 *
DepartmentResearch & Development               1.380e+01  1.615e+03   0.009 0.993182
DepartmentSales                               -8.055e-01  1.723e+03   0.000 0.999627
Education                                      3.803e-03  1.082e-01   0.035 0.971970
EducationFieldLife Sciences                   -9.740e-01  1.095e+00  -0.890 0.373712
EducationFieldMarketing                       -2.419e-01  1.159e+00  -0.209 0.834598
EducationFieldMedical                         -1.130e+00  1.091e+00  -1.036 0.300256
EducationFieldOther                           -1.376e+00  1.175e+00  -1.171 0.241654
EducationFieldTechnical Degree                 7.863e-02  1.130e+00   0.070 0.944529
EnvironmentSatisfaction                       -5.218e-01  1.028e-01  -5.077 3.83e-07 ***
GenderMale                                     1.367e-01  2.262e-01   0.604 0.545634
JobInvolvement                                -5.254e-01  1.520e-01  -3.457 0.000545 ***
JobLevel                                      -1.383e-01  2.687e-01  -0.515 0.606853
JobRoleHuman Resources                         1.506e+01  1.615e+03   0.009 0.992556
JobRoleLaboratory Technician                   1.288e+00  5.782e-01   2.228 0.025861 *
JobRoleManager                                 1.344e-01  9.792e-01   0.137 0.890793
JobRoleManufacturing Director                  3.697e-01  6.276e-01   0.589 0.555834
JobRoleResearch Director                      -1.172e+00  1.189e+00  -0.986 0.324261
JobRoleResearch Scientist                      7.077e-01  5.865e-01   1.207 0.227508
JobRoleSales Executive                         1.515e+01  6.005e+02   0.025 0.979870
JobRoleSales Representative                    1.663e+01  6.005e+02   0.028 0.977911
JobSatisfaction                               -3.116e-01  9.868e-02  -3.158 0.001590 **
MaritalStatusMarried                           4.129e-01  3.052e-01   1.353 0.176132
MaritalStatusSingle                            1.334e+00  3.152e-01   4.233 2.30e-05 ***
NumCompaniesWorked                             1.979e-01  5.321e-02   3.719 0.000200 ***
OverTimeYes                                    2.155e+00  2.399e-01   8.982  < 2e-16 ***
PercentSalaryHike                             -1.438e-02  2.990e-02  -0.481 0.630509
RelationshipSatisfaction                      -2.765e-01  1.026e-01  -2.694 0.007062 **
TrainingTimesLastYear                         -2.360e-01  8.849e-02  -2.667 0.007650 **
WorkLifeBalance                               -4.322e-01  1.547e-01  -2.793 0.005216 **
AgeGroup                                      -1.691e-01  8.368e-02  -2.021 0.043260 *
DistanceGroup                                  1.979e-01  6.751e-02   2.932 0.003371 **
YearsWithManagerGroup                          3.838e-01  1.608e-01   2.387 0.017004 *
AverageTenurePerJob_Group                     -3.602e-02  1.556e-01  -0.231 0.816955
YearsWithoutPromotion_WithCurrentManager_group -9.787e-01  1.815e-01  -5.393 6.95e-08 ***
TotalWorkingYears_Group                       -2.583e-01  1.487e-01  -1.737 0.082315 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 909.69  on 1029  degrees of freedom
Residual deviance: 575.87  on  992  degrees of freedom
AIC: 651.87

Number of Fisher Scoring iterations: 16

> qchisq(0.95, 992)
[1] 1066.385
```

Figure 4: Logistic Regression

A deviance table shows that as you add a variable there is a reduction in the deviance residual. However, some variables have a greater impact on reduction of deviance residual than others. The variables “OverTime”, “JobLevel”, “YearsWithoutPromotion_WithCurrentManager _group” and “BusinessTravel” have some of the greatest deviance residual reduction impact and each have low p-values, or p-values less than 0.05, signifying a significant influence on attrition. On the other hand, the variables “Education”, “YearsWithManagerGroup”, “AverageTenurePer

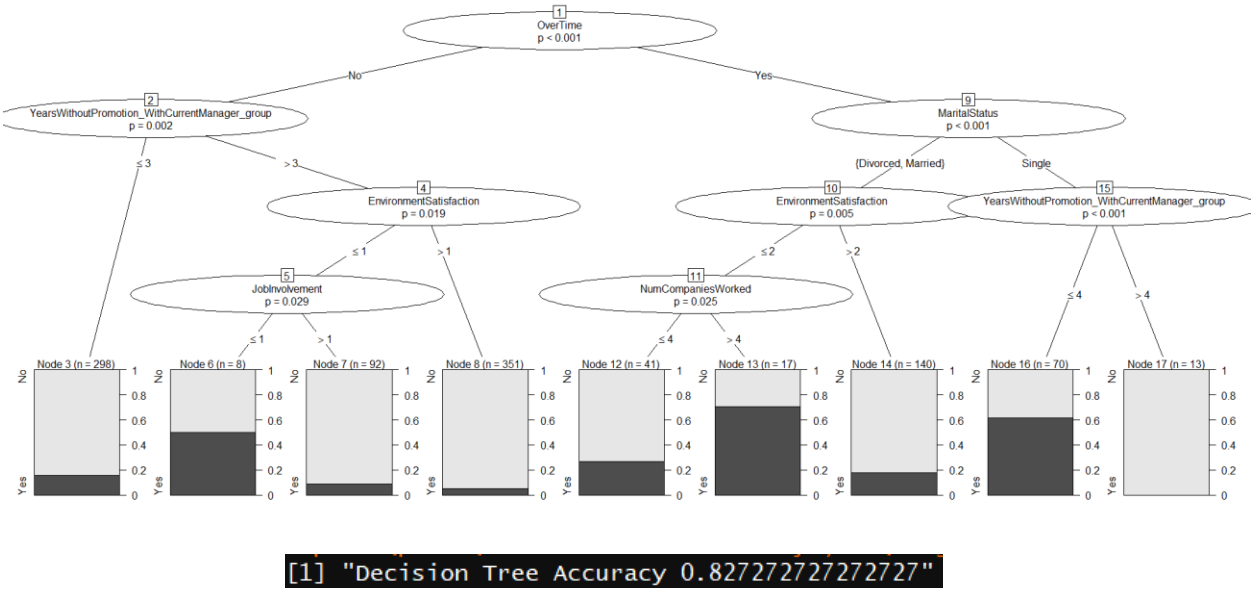
job_group", and "TotalWorkingYears_Group" have larger p-values and have a much smaller impact on the reduction of the residual deviance (Figure 5).

```
> anova(mod_fit, test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: Attrition

Terms added sequentially (first to last)


                                          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                                      1029     909.69
BusinessTravel                             2   20.042      1027     889.65 4.446e-05 ***
Department                                 2    6.262      1025     883.39 0.0436803 *
Education                                  1    0.955      1024     882.43 0.3284005
EducationField                             5    7.411      1019     875.02 0.1918219
EnvironmentSatisfaction                    1   12.845      1018     862.17 0.0003383 ***
Gender                                     1    0.084      1017     862.09 0.7713941
JobInvolvement                             1   13.973      1016     848.12 0.0001855 ***
JobLevel                                   1   61.379      1015     786.74 4.709e-15 ***
JobRole                                    8   13.768      1007     772.97 0.0880148 .
JobSatisfaction                            1    7.764      1006     765.21 0.0053308 **
MaritalStatus                              2   15.634      1004     749.57 0.0004028 ***
NumCompaniesWorked                         1   11.592      1003     737.98 0.0006625 ***
OverTime                                   1   86.254      1002     651.73 < 2.2e-16 ***
PercentSalaryHike                          1    0.077      1001     651.65 0.7810048
RelationshipSatisfaction                   1    6.615      1000     645.03 0.0101106 *
TrainingTimesLastYear                      1    7.924       999     637.11 0.0048790 **
WorkLifeBalance                            1    7.062       998     630.05 0.0078757 **
AgeGroup                                   1    8.287       997     621.76 0.0039934 **
DistanceGroup                              1    8.734       996     613.03 0.0031241 **
YearsWithManagerGroup                      1    1.630       995     611.40 0.2017557
AverageTenurePerJob_Group                  1    1.993       994     609.41 0.1580754
YearsWithoutPromotion_WithCurrentManager_group  1   30.452       993     578.95 3.421e-08 ***
TotalWorkingYears_Group                    1    3.081       992     575.87 0.0792097 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 5: Deviance Table

Decision Tree:

Though Logistic regression was found to be more accurate, it is often hard to interpret to the untrained eyw. Decision trees, on the other hand, are naturally easy to interpret. For this purpose, the decision tree model will also be used. The seven most significant variables from the logistic regression model were used to create the below decision tree (Figure 6). We can see that the accuracy is slightly lower than the logistic regression, approx. 5 percent difference. However, it is still admirable and worth exploring with its easily interpretive nature.




```
[1] "Decision Tree Accuracy 0.827272727272727"
```

Figure 6: Decision Tree

Of the seven variables, the most significant is Overtime. Meaning the variable Overtime and whether or not an employee has overtime is the best variable to indicate whether or not an employee will leave. Four other things we can pull from the decision tree are (1) Employees that are single and in group 4 or less (have 5 years or less) without a promotion with their current manager are more likely to leave. (2) If an employee has worked for more than four companies he/she is more like to leave than if they have worked for four or less companies. (3) If an employee has a job involvement level less than or equal to one he/she is more likely to leave. (4) Environmental Satisfaction seems to have and influence on whether an employee will leave a company especially if their environmental satisfaction level is less than or equal to one.

Discriminate Analysis:

To check if the data is discriminating two methods are used, Chi-squared and ROC Curve. The Chi-Squared was calculated above in the logistic regression section. As we saw, with a 95 percent confidence level, that the critical value was lager than the residual deviance. Thus, we

saw that the logistic regression model was discriminating and a good fit. As for the ROC curve, the area under the curve was calculated. A possible area value under the curve is between 0.50 and 1.00. If the curve for the selected model is 0.80 or greater, it can be determined that the model is discriminate. The calculated area under the curve for the logistic regression model was 0.806672. The logistic regression model was found to be discriminate.

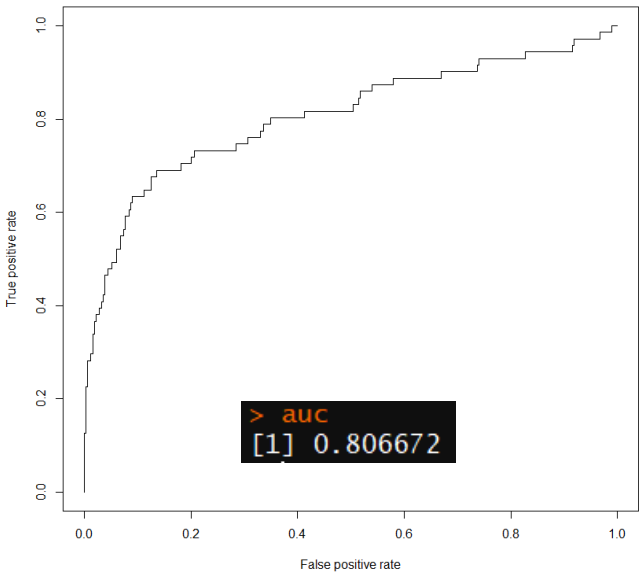


Figure 7: ROC Curve

Variable Importance – Logistic Regression Model:

From the caret package the varImp method was used to calculate/rank each variable in terms of importance/significance while predicting the variable "Attrition". We can see that OverTimeYes has the greatest impact on the variable attrition (Figure 8). Overtime in fact, had a 40 percent greater impact on Attrition than the next most significant variable YearsWithoutPromotion_withCurrent Manager_group. The variable OverTime has consistently been the most significant variable throughout the entire study independent of a particular model.

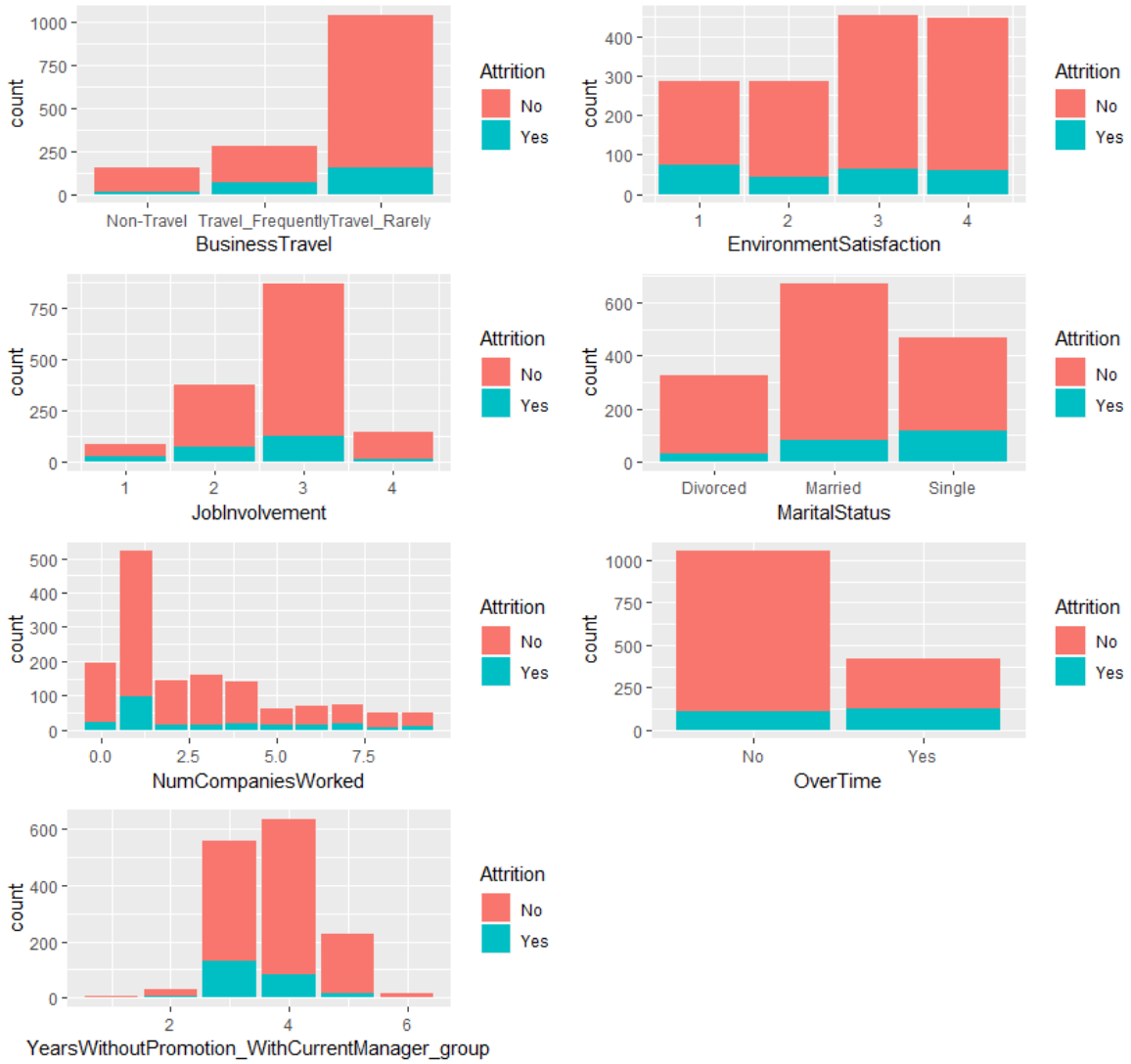
```
> varImp(sig_var)
glm variable importance

  only 20 most important variables shown (out of 35)

                                                Overall
OverTimeYes                                      100.00
YearsWithoutPromotion_WithCurrentManager_group    59.14
MaritalStatusSingle                               56.35
EnvironmentSatisfaction                           51.60
JobSatisfaction                                   50.22
JobInvolvement                                    45.03
BusinessTravelTravel_Frequently                   44.09
DistanceGroup                                     38.72
RelationshipSatisfaction                          29.36
WorkLifeBalance                                   28.93
NumCompaniesWorked                                28.76
`JobRoleLaboratory Technician`                    28.38
TrainingTimesLastYear                             26.54
BusinessTravelTravel_Rarely                       23.81
GenderMale                                        21.64
`JobRoleSales Representative`                     20.05
YearsWithManagerGroup                             18.33
JobLevel                                          17.34
AverageTenurePerJob_Group                         14.40
EducationFieldMedical                             14.07
```

Figure 8: Variable Importance

We started this research paper with the question: "What are the most likely reasons for an employee to leave a company and can the turnover possibility be predicted before it happens?" From the above we know that employees with overtime are more likely to leave. Yet, throughout the

analysis process we found that six other variables also have significance when predicting whether an employee will leave: BusinessTravel, EnvironmentSatisfaction, JobInvolvement, MaritalStatus, NumCompaniesWorked, YearsWithoutPromotion_WithCurrentManager_group. Each of these have p-values equal to or less than 0.001. Which signifies a significant influence to whether an employee will leave. Below is a recap analysis of the in 6 variables:

Figure: Significant Variables



There are many reasons why an employee might leave, but through these variables we start to have some insight. We begin to understand that employees that work overtime are more apt to leave, and employees that have worked at two or less companies are also more likely to leave. Furthermore, an employee that has not received a promotion within five years will leave

for another opportunity. It is also important to note that employees that are single will leave more often than someone that is married or divorced. The amount of business travel also seems to have an impact on attrition. Also, excellent working conditions or outstanding environmental satisfaction are important to employees. Employees that have a good clean environment to work in are more likely to stay.

The question now is, can attrition be predicted before it happens? Yes, I believe it can. Though this data set is an example, periodically employees are asked to perform surveys and end of year evaluations. These surveys and evaluations, coupled with HR records one could compile a data set that could be analyzed like unto our sample data set. You could then use the predictor variables to identify employees that are at risk for leaving. However, I would not recommend going to employees and saying, "It has come to our attention that you may consider leaving." Such an action would surely induce mistrust and discontent within the workplace. However, HR could use this information to make recommendations on policy changes, reassignments, promotions and other positive changes that would encourage employees to stay with the company thus decreasing the cost of attrition.

Proposed Action Summary & Limitations

The first action I would recommend to take is to identify the reasons for overtime and attempt to reduce if not eliminate the need for employees to work excessive hours. As overtime has one of the greatest impacts on attrition it would prudent to start there. A second recommendation would be to identify employees that have not received a promotion within three to ten years. These employees should be evaluated against their performance record and their managers recommendations for whether or not a promotion should be given.

It is important to note a major limitation to the analysis. Though we can take a data set and perform analysis and attempt to discover why an employee may leave, the reasons could be

endless and go far beyond the variables contained therein. For example, though we know that those that work overtime are more likely to leave, we do not know if it is merely the extra hours or whether it is something else entirely. Perhaps the environmental satisfaction is low and the employee would be happy to work a few extra hours if their environment was so not so poor. To predict human emotion is very difficult if not impossible.

To overcome this limitation, I purpose two approaches to further study the data set. First, surveys can be sent out to employees that have previously left the company. The surveys could ask for further information and understanding on the reasons they left. Questions could include but are not limited to the following: Did you feel valued at work? If there was another position that interested you at (company name) would you return? Why or why not? Out of a scale from 1 to 10, one being unsatisfactory and ten being Outstandingly satisfactory, rate your experience with (company name). Out of on a scale of 1 through 5 rate your satisfaction with your manager. etc. There are many options for questions but they should seek to gain further understanding. Also, though selective answer questions are good, open-ended questions should also be used to give a chance for the former employee to express their options without being confined to a predetermined set of answers.

The second approach as mentioned previously, HR should use the analysis to direct decisions. The purpose of the analysis is not to make decisions but to help in the guiding of decisions. Furthermore, the HR department should focus on a follow-up survey for current employees that focuses on deepening the understanding of the analysis. The goal being that it would shine a light on policies and procedures that could be changed to better the workplace experience and environment.

Expected Benefits

Though there are an infinite amount of plausible benefits to this study, I will mention just three that I find to be the most significate. It is somewhat obvious, but nonetheless important, that the first benefit to applying and furthering this study would be reduced attrition/turnover of employees. In the beginning of this summary we surmised that the cost of attrition could be 1.5 to 2.0 times an employee's annual salary (Altman, 2017). This can be an enormous cost for both big and small companies alike. No one wants to spend the time and money required to train a new an employee without reaping the benefits of that commitment.

The second benefit, and somewhat of a segue from the first benefit, is increased revenue. Of course, by pinpointing the sore points in HR and lowering the turnover rate the company can save an untold amount in onboarding and employee replacement costs. However, not having to replace an employee is not the only way it will increase revenue. It takes up to 2 years for an employee to become fully productive (Marsden-Huggins, 2017). By having employees with longer tenure, you will have more years of greater productivity thus a plausible increase in revenue.

The third benefit that I will mention would be increased job satisfaction which in turn would make a happier work place which would in turn increase revenue. "A recent study by economists at the University of Warwick found that happiness led to a 12% spike in productivity, while unhappy workers proved 10% less productive. As the research team put it, "We find that human happiness has large and positive causal effects on productivity. Positive emotions appear to invigorate human beings. (Revesencio, 2015)" The key to more positive employees is to invest in employee support and satisfaction.

Though it would be hard if not impossible to calculate a specific percentage increase or decrease within these three benefits that could be applied to each company there is one thing that

we can be sure of, when you take the steps to improve working conditions and invest in employee support and satisfaction you will have happier workers and a greater potential for a return on your investment.

References

Altman, J. (2017, 01 18). *How Much Does Employee Turnover Really Cost?* Retrieved from HUFFPOST: https://www.huffingtonpost.com/entry/how-much-does-employee-turnover-really-cost_us_587fbaf9e4b0474ad4874fb7

Marsden-Huggins, S. (2017, May 1). *How Long Does it Take An Employee to Be Fully Productive?* . Retrieved from RecruitShop : https://recruitshop.com.au/long-take-employee-fully-productive/

Revesencio, J. (2015, July 22). *Why Happy Employees Are 12% More Productive*. Retrieved from Fast Company: https://www.fastcompany.com/3048751/happy-employees-are-12-more-productive-at-work