Business Data Analytics Final Project

BSAD 443 - Professor Osman S. Abraz

John Chrusciel Jr & Matthew Salmon

29 October 2018

### 1. Executive Summary

Our services were engaged to solve the employee turnover problem among the seemingly satisfied and well-evaluated, productive employees. We analyzed extensive data regarding the state of the firms human resources. The data included quantitative and qualitative data analyzed according to the strictest data guidelines of the emerging field of data science. This analysis was guided by our deep domain knowledge and unique insights into the company gleaned while performing the duties set forth in our scope of work agreement.

Our analysis uncovered several areas for further exploration to with be believe your safety and plant security (EHSS) plan needs to be reviewed based on the data we received, you may want to revisit your data governance policies and procedures. Most pressing however, is your need to work with your human resources department to better balance workflows while retaining the high productivity you are accustomed to. We believe well balanced workflows will lead to better satisfaction levels which will stave off employee flight. Additionally, your compensation package needs to be looked at against your industry competitors. Finally, your employees need an engagement program that will give them a renewed optimism about their chances for advancement and recognition.

### 2. Business Understanding

Our firm was engaged by a company to study their employee turnover problem. We intended to take an empirical approach to determining the root cause of this turnover problem.

#### 2.1 Data Sizing & Description

This analysis is based entirely on the data we received from the entity engaging our academic/consulting services. Although every reasonable and prudent attempt has been made to ensure that commonly accepted and sound business data analytics principles have been adhered to during the performance of the work covered under the scope of our agreement, we must add the caveat that the results and recommendations we are making are based on data provided to us by your organization.

We were provided with two data sets, the training data set consisting of the HR records of 10,000 current and separated employees and a smaller testing data set composed of the records of 5,000 other different current and separated employees which we assume was not included in the training data set. We are also operating under the assumption that the data sets were properly sampled and partitioned.

Proceding forward with the understanding that both data sets were generated by the sponsoring organization without our technical oversight, the ultimate validity of the analysis, assessments, and recommendations we have formed under the scope of our engagement must be intrerpreted of light of this risk.

Both the Training Data Set (TRDS) and the Testing Data Set (TEDS) were received electronically via an MS Excel Workbook attached to an email from the sponsor’s HR Department. The workbook (HR.xsls) consisted of two tabbed pages, one tab named “Training”" contained the with the training data, the other tab labelled “Testing” which contained the testing data.

File HR.xlsx was created by the sponsor on 04/16/2017 at 5:06 PM and was last saved by our sponsor on 10/10/2018 at 5:30 PM. The file was transmitted to and received by our organization with a size of 668 KB on 10/26/2018. The content type is application/vnd.openxmlformats-officedocument.spreadsheetml.sheet.

The data is structured and is of mixed type, both numerical and categorical. The information contained is proprietary and can be considered confidential.

The TRDS, as received, in it’s native Microsoft Excel format is 10,001 rows long including the header row. Each row has 10 attributes. The TEDS is similarly structured except that it is 5,000 rows long.

Both datasets are structured exactly the same. The attributes (columns) are as follows:

**satisfaction\_level:** This is Categoric Ordinal Data ranked from 0 to 1. A value approaching 1 is associated with high job satisfaction and a value nearing zero can be associated with low job satisfaction.

We are assuming that this is the job satisfaction value was self-reported during the employees most recent formal review. For those employees separated from service, we do not know if this score was recorded from information disseminated during their normal assessment review or if it was recorded as a result of their exit interview.

We would like to have also seen the history of the employees satisfaction levels as recorded throughout their tenure as taken and recorded during the regular review process.

**last\_evaluation:** Categoric Ordinal Data ranked from 0 to 1. A value approaching one is a favorable review, a value approaching zero is unfavorable. We assume this value is the manager reported.

It might have been helpful to see the values over time.

**number\_project:** This is a Numeric Discrete Data. It is the count of projects an employee is currently associated with. We are making the assumption that this is the current number of projects a given employee is tasked with.

It might have been helpful to see the roles the employee played in the projects the were assigned to. Were they the PM, the team leader, or were they relegated to a support role.

**average\_monthly:** This is a Numerical Continuous Data. The numerical value representing the number of hours an employee averages per month. We do not know the basis for average computation. Is this the current rolling one month average? Is this the employees total average weekly hours since the employee was hired?

It would have like to have seen several averages, a total average hours since hire, a rolling annual average, a rolling quarterly average and a rolling monthly average.

**time\_spend\_company:** This is a Numerical Discrete Data and represents the number of the full years that an employee has worked at the company.

**work\_accident:** This is Categorical Nominal Data and only tells us if an employee had a work related accident or not. It does not track the number of accidents nor does it track the severity of accidents.

**left:** This Categorical Nominal Data, a label. 0 is the code for current with firm, 1 is separated from service.

**promotion\_last-5years:** Categorical Nominal Data. 0 labels no promotion, 1 represents a received a promotion.

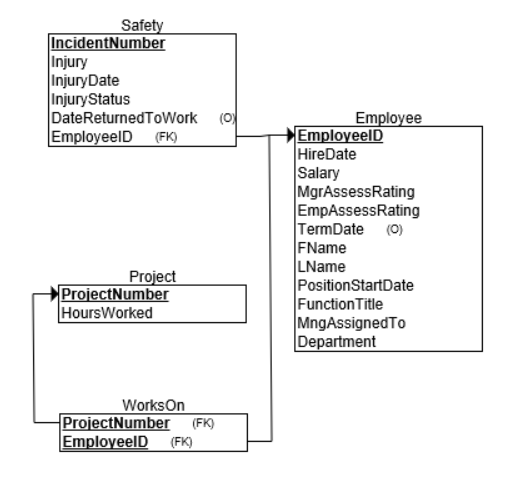
**department:** This is Categorical Data expressed textually. It is nominal in nature. We will need to transform this data. Overtly, this value indicates which deparment the employee is a member of.

Role within department might have been good data to see as well.

**salary:** Categorical Data expresssed in text that is Ordinal in nature. High, Medium, Low. Depending on the method used, we are probably going to have to transform this data.

I would have like to have seen information on raises and frequency of raises as well.

#### 2.2 Data Understanding, Analysis & Preparation



1. This is we think the ideal warehouse would look like that developed the ideal predictive models, but this is not what we are working with.

Now that we have a thorough understanding of the data that we were given, let us load the training set and explore it.

1. Now that we have understand the data, the raw data needed to be transformed. We converted, cleaned, and derived data. The process included removing duplicates and checking for outliers. While preparing the data, we discovered one employee working 350 hours a month, meaning 80 hours a week. We also looked at the business logic and found an issue. Although this may be true and the record was kept in the data, we find it hard to believe this average is probable.
2. Preprocess included scanning the data, cleaning (removing duplicates and outliers, and transforming the data into a form usable by our statistical methods. This includes converting numeric categorical data into factor data for the sake of logistical regression.)
3. Data governance needs to be addressed to prevent 350 hour month as we addressed above. Data governance policy needs to be established to ensure integrity/veracity of data that will be used for analysis.
4. Suggest hard coding a maximum allowable, enterable value of 303 for monthly hours with manual managerial override if more hours must be worked by one individual. This will let us properly treat outliers.
5. All the data have been qualitatively and quantitatively described below.

train <- read.csv("training.csv", header = TRUE, sep = ",")

str(train)

## 'data.frame': 9999 obs. of 10 variables:  
## $ satisfaction\_level : num 0.38 0.8 0.11 0.72 0.37 0.41 0.1 0.92 0.89 0.42 ...  
## $ last\_evaluation : num 0.53 0.86 0.88 0.87 0.52 0.5 0.77 0.85 1 0.53 ...  
## $ number\_project : int 2 5 7 5 2 2 6 5 5 2 ...  
## $ average\_montly\_hours : int 157 262 272 223 159 153 247 259 224 142 ...  
## $ time\_spend\_company : int 3 6 4 5 3 3 4 5 5 3 ...  
## $ Work\_accident : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ left : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ promotion\_last\_5years: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ department : Factor w/ 10 levels "accounting","hr",..: 8 8 8 8 8 8 8 8 8 8 ...  
## $ salary : Factor w/ 3 levels "high","low","medium": 2 3 3 2 2 2 2 2 2 2 ...

train[,3] <-as.factor(train[,3])  
train[,5] <-as.factor(train[,5])  
train[,6] <-as.factor(train[,6])  
train[,7] <-as.factor(train[,7])  
train[,8] <-as.factor(train[,8])  
str(train)

## 'data.frame': 9999 obs. of 10 variables:  
## $ satisfaction\_level : num 0.38 0.8 0.11 0.72 0.37 0.41 0.1 0.92 0.89 0.42 ...  
## $ last\_evaluation : num 0.53 0.86 0.88 0.87 0.52 0.5 0.77 0.85 1 0.53 ...  
## $ number\_project : Factor w/ 6 levels "2","3","4","5",..: 1 4 6 4 1 1 5 4 4 1 ...  
## $ average\_montly\_hours : int 157 262 272 223 159 153 247 259 224 142 ...  
## $ time\_spend\_company : Factor w/ 5 levels "2","3","4","5",..: 2 5 3 4 2 2 3 4 4 2 ...  
## $ Work\_accident : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## $ left : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...  
## $ promotion\_last\_5years: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## $ department : Factor w/ 10 levels "accounting","hr",..: 8 8 8 8 8 8 8 8 8 8 ...  
## $ salary : Factor w/ 3 levels "high","low","medium": 2 3 3 2 2 2 2 2 2 2 ...

summary(train)

## satisfaction\_level last\_evaluation number\_project average\_montly\_hours  
## Min. :0.0900 Min. :0.3600 2:1453 Min. : 96.0   
## 1st Qu.:0.4500 1st Qu.:0.5600 3:2826 1st Qu.:157.0   
## Median :0.6500 Median :0.7200 4:2980 Median :200.0   
## Mean :0.6226 Mean :0.7175 5:1861 Mean :200.7   
## 3rd Qu.:0.8200 3rd Qu.:0.8700 6: 734 3rd Qu.:244.0   
## Max. :1.0000 Max. :1.0000 7: 145 Max. :350.0   
##   
## time\_spend\_company Work\_accident left promotion\_last\_5years  
## 2:2451 0:8496 0:7999 0:9901   
## 3:4394 1:1503 1:2000 1: 98   
## 4:1740   
## 5: 972   
## 6: 442   
##   
##   
## department salary   
## sales :2648 high : 724   
## technical:1964 low :4915   
## support :1571 medium:4360   
## IT : 796   
## RandD : 608   
## marketing: 548   
## (Other) :1864

The average satisfaction level is about 62%

The average employee is assigned to 4 projects and the avg employee works 200 hours per month, 46.15 hours per week.

72.6% of the employees have 4 years of tenure or less.

The OSHA national AVG for accidents is 2.9% of all industries (heavy and light). Of 10,000 employees at this firm, the accident rate is over 15%. This is indicative of a frenetic, taxing, safety-lax, low morale environment.

This unsafe environment offers little hope for advancement as the promotion rate over five years has been less than 1%.

Looking at departments, sales, technical, and support are the majority of the employees and 93% of the employees are paid either low or medium level salary.

library("Hmisc")

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

##   
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':  
##   
## format.pval, units

# separate the first 2000 observations (those who left)from the rest of the data

separated <- train[ which(train$left == 1), ]

str(separated)

## 'data.frame': 2000 obs. of 10 variables:  
## $ satisfaction\_level : num 0.38 0.8 0.11 0.72 0.37 0.41 0.1 0.92 0.89 0.42 ...  
## $ last\_evaluation : num 0.53 0.86 0.88 0.87 0.52 0.5 0.77 0.85 1 0.53 ...  
## $ number\_project : Factor w/ 6 levels "2","3","4","5",..: 1 4 6 4 1 1 5 4 4 1 ...  
## $ average\_montly\_hours : int 157 262 272 223 159 153 247 259 224 142 ...  
## $ time\_spend\_company : Factor w/ 5 levels "2","3","4","5",..: 2 5 3 4 2 2 3 4 4 2 ...  
## $ Work\_accident : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## $ left : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...  
## $ promotion\_last\_5years: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## $ department : Factor w/ 10 levels "accounting","hr",..: 8 8 8 8 8 8 8 8 8 8 ...  
## $ salary : Factor w/ 3 levels "high","low","medium": 2 3 3 2 2 2 2 2 2 2 ...

summary(separated)

## satisfaction\_level last\_evaluation number\_project average\_montly\_hours  
## Min. :0.09 Min. :0.450 2:865 Min. :126.0   
## 1st Qu.:0.11 1st Qu.:0.520 3: 38 1st Qu.:146.0   
## Median :0.41 Median :0.790 4:237 Median :225.0   
## Mean :0.44 Mean :0.721 5:343 Mean :207.9   
## 3rd Qu.:0.73 3rd Qu.:0.900 6:372 3rd Qu.:262.0   
## Max. :0.92 Max. :1.000 7:145 Max. :350.0   
##   
## time\_spend\_company Work\_accident left promotion\_last\_5years  
## 2: 31 0:1895 0: 0 0:1992   
## 3:882 1: 105 1:2000 1: 8   
## 4:496   
## 5:482   
## 6:109   
##   
##   
## department salary   
## sales :556 high : 48   
## technical:390 low :1182   
## support :312 medium: 770   
## IT :159   
## hr :113   
## marketing:112   
## (Other) :358

testing <- read.csv("TESTING.csv", header = TRUE, sep = ",")

str(testing)

## 'data.frame': 5000 obs. of 10 variables:  
## $ satisfaction\_level : num 0.26 0.93 0.26 0.9 0.16 0.65 0.62 0.66 0.13 0.96 ...  
## $ last\_evaluation : num 0.91 0.49 0.61 0.63 0.6 0.63 0.64 0.51 0.66 0.92 ...  
## $ number\_project : int 5 4 3 4 6 3 2 3 4 3 ...  
## $ average\_montly\_hours : int 163 255 213 163 246 146 236 165 201 150 ...  
## $ time\_spend\_company : int 6 2 6 3 4 4 3 3 5 2 ...  
## $ Work\_accident : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ left : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ promotion\_last\_5years: int 0 0 0 0 0 0 0 0 0 0 ...  
## $ department : Factor w/ 10 levels "accounting","hr",..: 3 4 8 1 9 10 4 5 5 5 ...  
## $ salary : Factor w/ 3 levels "high","low","medium": 3 1 2 3 3 3 3 3 3 3 ...

testing[,3] <-as.factor(testing[,3])  
testing[,5] <-as.factor(testing[,5])  
testing[,6] <-as.factor(testing[,6])  
testing[,7] <-as.factor(testing[,7])  
testing[,8] <-as.factor(testing[,8])  
str(testing)

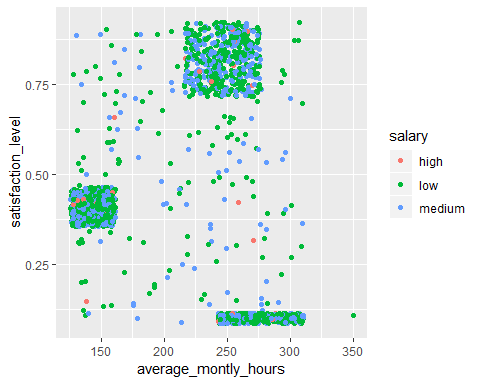
## 'data.frame': 5000 obs. of 10 variables:  
## $ satisfaction\_level : num 0.26 0.93 0.26 0.9 0.16 0.65 0.62 0.66 0.13 0.96 ...  
## $ last\_evaluation : num 0.91 0.49 0.61 0.63 0.6 0.63 0.64 0.51 0.66 0.92 ...  
## $ number\_project : Factor w/ 5 levels "2","3","4","5",..: 4 3 2 3 5 2 1 2 3 2 ...  
## $ average\_montly\_hours : int 163 255 213 163 246 146 236 165 201 150 ...  
## $ time\_spend\_company : Factor w/ 5 levels "2","3","4","5",..: 5 1 5 2 3 3 2 2 4 1 ...  
## $ Work\_accident : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## $ left : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## $ promotion\_last\_5years: Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...  
## $ department : Factor w/ 10 levels "accounting","hr",..: 3 4 8 1 9 10 4 5 5 5 ...  
## $ salary : Factor w/ 3 levels "high","low","medium": 3 1 2 3 3 3 3 3 3 3 ...

summary(testing)

## satisfaction\_level last\_evaluation number\_project average\_montly\_hours  
## Min. :0.090 Min. :0.3600 2: 935 Min. : 16.0   
## 1st Qu.:0.410 1st Qu.:0.5500 3:1229 1st Qu.:154.0   
## Median :0.600 Median :0.6300 4:1385 Median :196.0   
## Mean :0.548 Mean :0.6586 5: 900 Mean :197.5   
## 3rd Qu.:0.660 3rd Qu.:0.6600 6: 551 3rd Qu.:246.0   
## Max. :1.000 Max. :1.0000 Max. :309.0   
##   
## time\_spend\_company Work\_accident left promotion\_last\_5years  
## 2: 793 0:4334 0:3429 0:4779   
## 3:2049 1: 666 1:1571 1: 221   
## 4: 817   
## 5: 501   
## 6: 840   
##   
##   
## department salary   
## sales :1492 high : 513   
## technical : 756 low :2401   
## support : 658 medium:2086   
## IT : 431   
## product\_mng: 372   
## management : 366   
## (Other) : 925

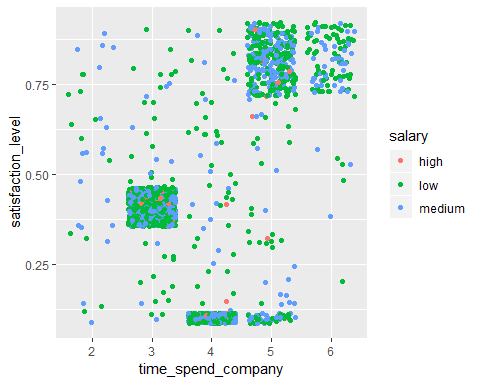
ggplot(separated, aes(x = average\_montly\_hours, y =satisfaction\_level)) + geom\_jitter(aes(color = salary))

Data Visualization

 pay plays an issue…overworked underpaid.

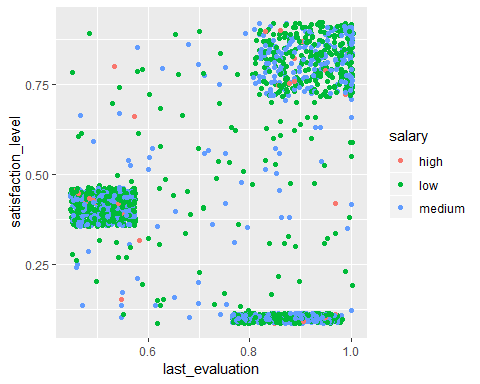
Analyzing only the subset of employees that are no longer with the firm, we find that they cluster into 3 distinct groups. The first group consists of low satisfaction, high average hours, and low pay. The next group consists of moderately dissatisfied, low wage workers with working a relatively low amount of hours per month. The third group consists of satisfied employees who are working more than average monthly hours who are earning low and medium salary.

ggplot(separated, aes(x = time\_spend\_company, y =satisfaction\_level)) + geom\_jitter(aes(color = salary))



Analyzing this chart we see nearly all of employees who have left have tenure longer than 68% of their peers. This is beyond 3 years of tenure.

ggplot(separated, aes(x = last\_evaluation, y =satisfaction\_level)) + geom\_jitter(aes(color = salary))



### This chart is the most puzzling. In the upper right quadrant we see people who scored well on their last evaluation and have expressed high satisfaction level are leaving. These employees have left. Interesting to note that the upper right quadrant cluster represents separated employees shows a mix of medium to low wage workers. Why are people with medium wages and are happy leaving the company?

### 3. Modeling & Solving

Here we will begin modeling for a solution to the problem. We are starting with a stepped linear model and with salary data removed.

mod\_sep <- separated[, -(7:10) ]  
mod <- lm(satisfaction\_level ~., data = mod\_sep)  
  
step(mod)

## Start: AIC=-8889.16  
## satisfaction\_level ~ last\_evaluation + number\_project + average\_montly\_hours +   
## time\_spend\_company + Work\_accident  
##   
## Df Sum of Sq RSS AIC  
## - Work\_accident 1 0.0009 23.182 -8891.1  
## <none> 23.181 -8889.2  
## - average\_montly\_hours 1 0.3991 23.580 -8857.0  
## - last\_evaluation 1 1.0788 24.260 -8800.2  
## - time\_spend\_company 4 6.7921 29.973 -8383.2  
## - number\_project 5 19.7845 42.965 -7665.0  
##   
## Step: AIC=-8891.08  
## satisfaction\_level ~ last\_evaluation + number\_project + average\_montly\_hours +   
## time\_spend\_company  
##   
## Df Sum of Sq RSS AIC  
## <none> 23.182 -8891.1  
## - average\_montly\_hours 1 0.3982 23.580 -8859.0  
## - last\_evaluation 1 1.0792 24.261 -8802.1  
## - time\_spend\_company 4 6.8057 29.987 -8384.3  
## - number\_project 5 19.7943 42.976 -7666.5

##   
## Call:  
## lm(formula = satisfaction\_level ~ last\_evaluation + number\_project +   
## average\_montly\_hours + time\_spend\_company, data = mod\_sep)  
##   
## Coefficients:  
## (Intercept) last\_evaluation number\_project3   
## 0.3969592 0.2990294 0.0919622   
## number\_project4 number\_project5 number\_project6   
## 0.1372687 0.1233390 -0.2824668   
## number\_project7 average\_montly\_hours time\_spend\_company3   
## -0.2979810 -0.0005686 -0.0578488   
## time\_spend\_company4 time\_spend\_company5 time\_spend\_company6   
## -0.1196544 0.1144793 0.1502143

mod <- lm(satisfaction\_level ~ last\_evaluation + number\_project + average\_montly\_hours + time\_spend\_company, data = separated)

model <-lm(formula = satisfaction\_level ~ last\_evaluation + number\_project +   
 average\_montly\_hours + time\_spend\_company, data = mod\_sep)

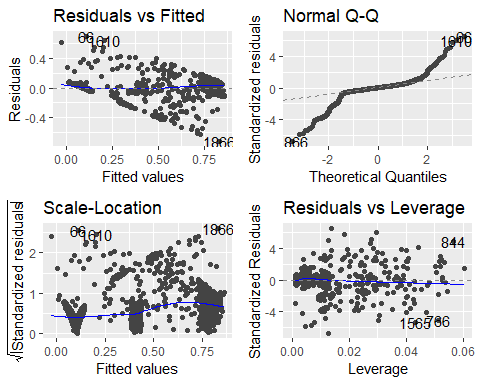
The best linear calculated by the computer has an AIC (Akaine index) of 8891.08 appears to be high.

A linear model might not be the ideal tool. There might be better tools than the linear model for this particular dataset.

summary(model)

##   
## Call:  
## lm(formula = satisfaction\_level ~ last\_evaluation + number\_project +   
## average\_montly\_hours + time\_spend\_company, data = mod\_sep)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.72328 -0.02711 0.00557 0.03551 0.69738   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.970e-01 3.259e-02 12.179 < 2e-16 \*\*\*  
## last\_evaluation 2.990e-01 3.108e-02 9.620 < 2e-16 \*\*\*  
## number\_project3 9.196e-02 2.212e-02 4.158 3.34e-05 \*\*\*  
## number\_project4 1.373e-01 1.742e-02 7.878 5.42e-15 \*\*\*  
## number\_project5 1.233e-01 1.753e-02 7.035 2.74e-12 \*\*\*  
## number\_project6 -2.825e-01 1.819e-02 -15.530 < 2e-16 \*\*\*  
## number\_project7 -2.980e-01 1.970e-02 -15.128 < 2e-16 \*\*\*  
## average\_montly\_hours -5.686e-04 9.729e-05 -5.844 5.94e-09 \*\*\*  
## time\_spend\_company3 -5.785e-02 2.319e-02 -2.495 0.0127 \*   
## time\_spend\_company4 -1.197e-01 2.153e-02 -5.557 3.12e-08 \*\*\*  
## time\_spend\_company5 1.145e-01 2.098e-02 5.457 5.46e-08 \*\*\*  
## time\_spend\_company6 1.502e-01 2.300e-02 6.531 8.29e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.108 on 1988 degrees of freedom  
## Multiple R-squared: 0.8345, Adjusted R-squared: 0.8336   
## F-statistic: 911.5 on 11 and 1988 DF, p-value: < 2.2e-16

library(ggfortify)  
autoplot(model)



# lets try a logistic regression model

library(MASS)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following object is masked from 'package:MASS':  
##   
## select

## The following objects are masked from 'package:Hmisc':  
##   
## src, summarize

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

model <- glm(left ~., data = train, family = binomial) %>% stepAIC(trace=FALSE)  
summary(model)

##   
## Call:  
## glm(formula = left ~ satisfaction\_level + last\_evaluation + number\_project +   
## average\_montly\_hours + time\_spend\_company + Work\_accident +   
## salary, family = binomial, data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.9191 -0.3557 -0.1368 -0.0324 3.9835   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -5.288e+00 3.488e-01 -15.162 < 2e-16 \*\*\*  
## satisfaction\_level -2.023e+00 1.683e-01 -12.023 < 2e-16 \*\*\*  
## last\_evaluation 2.408e+00 2.600e-01 9.262 < 2e-16 \*\*\*  
## number\_project3 -5.324e+00 1.917e-01 -27.773 < 2e-16 \*\*\*  
## number\_project4 -3.846e+00 1.271e-01 -30.261 < 2e-16 \*\*\*  
## number\_project5 -3.194e+00 1.272e-01 -25.104 < 2e-16 \*\*\*  
## number\_project6 -2.382e+00 1.506e-01 -15.818 < 2e-16 \*\*\*  
## number\_project7 1.449e+01 1.768e+02 0.082 0.935   
## average\_montly\_hours 8.303e-03 8.823e-04 9.411 < 2e-16 \*\*\*  
## time\_spend\_company3 2.885e+00 2.092e-01 13.790 < 2e-16 \*\*\*  
## time\_spend\_company4 2.877e+00 2.254e-01 12.764 < 2e-16 \*\*\*  
## time\_spend\_company5 4.742e+00 2.225e-01 21.311 < 2e-16 \*\*\*  
## time\_spend\_company6 3.451e+00 2.444e-01 14.120 < 2e-16 \*\*\*  
## Work\_accident1 -1.535e+00 1.315e-01 -11.668 < 2e-16 \*\*\*  
## salarylow 1.672e+00 1.902e-01 8.790 < 2e-16 \*\*\*  
## salarymedium 1.339e+00 1.921e-01 6.969 3.18e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 10007.6 on 9998 degrees of freedom  
## Residual deviance: 5130.2 on 9983 degrees of freedom  
## AIC: 5162.2  
##   
## Number of Fisher Scoring iterations: 15

prmodel2 <- glm(left ~ satisfaction\_level + last\_evaluation + number\_project + average\_montly\_hours + time\_spend\_company + salary, family = binomial, data = train)

Here we try a stepped linear model that seems to have a better AIC score.

summary(prmodel2)

##   
## Call:  
## glm(formula = left ~ satisfaction\_level + last\_evaluation + number\_project +   
## average\_montly\_hours + time\_spend\_company + salary, family = binomial,   
## data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.8462 -0.3671 -0.1435 -0.0348 4.0231   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -5.354e+00 3.451e-01 -15.512 < 2e-16 \*\*\*  
## satisfaction\_level -2.004e+00 1.650e-01 -12.140 < 2e-16 \*\*\*  
## last\_evaluation 2.365e+00 2.561e-01 9.234 < 2e-16 \*\*\*  
## number\_project3 -5.311e+00 1.907e-01 -27.847 < 2e-16 \*\*\*  
## number\_project4 -3.822e+00 1.254e-01 -30.481 < 2e-16 \*\*\*  
## number\_project5 -3.157e+00 1.253e-01 -25.191 < 2e-16 \*\*\*  
## number\_project6 -2.389e+00 1.490e-01 -16.033 < 2e-16 \*\*\*  
## number\_project7 1.465e+01 1.778e+02 0.082 0.934   
## average\_montly\_hours 8.343e-03 8.696e-04 9.594 < 2e-16 \*\*\*  
## time\_spend\_company3 2.899e+00 2.078e-01 13.954 < 2e-16 \*\*\*  
## time\_spend\_company4 2.896e+00 2.238e-01 12.942 < 2e-16 \*\*\*  
## time\_spend\_company5 4.708e+00 2.207e-01 21.331 < 2e-16 \*\*\*  
## time\_spend\_company6 3.420e+00 2.440e-01 14.017 < 2e-16 \*\*\*  
## salarylow 1.576e+00 1.876e-01 8.398 < 2e-16 \*\*\*  
## salarymedium 1.244e+00 1.896e-01 6.561 5.34e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 10007.6 on 9998 degrees of freedom  
## Residual deviance: 5292.8 on 9984 degrees of freedom  
## AIC: 5322.8  
##   
## Number of Fisher Scoring iterations: 15

#### 3.1 Model Selection

Based on our analysis so far, because of the type of data we were given to analyze, mainly categorical data, we believe a logistic model will give us the best predictive values.

#### 3.2 Demonstrating Value & Accuracy

Let us test the final logistic regression model and measure the misclassification error by using a confusion matrix and then tabulating the results..

p1 <- predict(prmodel2, train, type = "response")  
head(p1)

## 1 2 3 4 5 6   
## 0.7157611 0.2260772 1.0000000 0.5618168 0.7184134 0.6811587

pred1 <- ifelse(p1 > .05, 1, 0)  
tab1 <- table(predicted =pred1, Actual = train$left)  
tab1

## Actual  
## predicted 0 1  
## 0 5079 48  
## 1 2920 1952

misclass <- (1-sum(diag(tab1))/sum(tab1))  
misclass

## [1] 0.2968297

p2 <- predict(prmodel2, testing, type ="response")  
pred2 <- ifelse(p2 > .05, 1, 0)  
tab2 <- table(predicted = pred2, testing = testing$left)  
tab2

## testing  
## predicted 0 1  
## 0 2167 42  
## 1 1262 1529

misclass <- (1-sum(diag(tab2))/sum(tab2))  
misclass

## [1] 0.2608

### Our logistic model has a 30% classification error when run against the training set. This error drops to 26% when tested against actual value.

### Our predicted values when tested against the training set generated a 30% misclassification error. When our predicted values were compared against the test set, our misclassification error drops to 26%. We believe that our model accurately predicts turnover at the 70% confidence level.

### 4. Reporting and Communicating Results

The company is losing a mix of good and bad employees. A company should be focused on how to retain their best talent, but they first need to know what is important to their people. They have engaged our services, but we are given data that is NOT informative.

To properly predict employee turnover there needs to be relevant data about employee performance and job status. There is little more than tenure rounded to whole years data to work with, but nothing of their actual functions or workflow that could be leading to their turnover rate. Becoming as granular to record who hired and interviewed the employees could lead to the employer discovering that the incorrect talent is being recruited by a sing individual or that proper background assessment is not being performed.

We suggest studying salary levels for positions at the firm relative to what the industry, as a whole, pays for the same position. Obviously this is a high pressure environment. Employees work long hours, are assigned to multiple projects and it appears that lower and middle-salaried employees are fleeing. The job market is blazing hot right now and we know it is cheaper to retain an employee than to attract a new employee. Perhaps the human resources department should work on safety, employee engagement in the form of promotion and recognition along with revamping their current pay structure.