Employee Attrition Model

Authors: Levin Reichmuth, Jorit Studer and Taejun Moon

Module: Machine Learning I

Submitted on June 10th, 2022

Supervisor: Dr. Matteo Tanadini, Daniel Meister and Dr. Alberto Paganini

Lucerne University of Applied Sciences and Arts

Table of Contents

1 Introduction

Employees, according to Swaminathan & Hagarty (2020), are the foundation of any business. Its success is largely determined by the quality of its employees and their ability to stay with the company. Organizations confront a number of issues as a result of staff attrition:

1. Training new personnel is costly in terms of both money and time.
2. Potential to lose experienced employees
3. Productivity impact
4. Profitability impact

Therefore, IBM data scientists created a fictitious data set as a challenge for data scientists. Among the data types are metrics such as education level, job satisfaction, and commute distance. The data set can be found on the company's GitHub account (IBM, 2019).

2 Methodology

The following topics are layed out through out this paper:

1. Exploration: Lead: Jorit Studer and Levin Reichmuth
2. Linear Models and Non-linearity (GAM), Lead: Levin Reichmuth
3. Generalised Linear Models (GLM), Lead: Jorit Studer
4. SVM and Neural Networks, Lead: Taejun Moon
5. Optimization, Lead: Jorit Studer and Taejun Moon

3 Data preparation

3.1 Data Transformation and Sanity Check

The code for this part is left out from the PDF due to its length...

3.2 Data Cleaning

```
emp_attrition <- emp_attrition %>% dplyr::select(-c(EmployeeCount, StandardHours, Over18))
```

- EmployeeCount (represents the head count which is 1 for all employee, hence drop this)
- StandardHours (StandardHours for all employee's is 80, therefore this data has a 9/80 work schedule. Hence, employees work 80 hours in 9 days. So not a standard as 5/42 as in switzerland, we drop this)
- Over18 (all employee's are 18 or above and it's capured in age, hence drop this variable)

3.3 Missing Value Check

```
# Do we have any missing values?
sapply(emp_attrition, function(x) all(is.na(x) | x == '' ))
```

There are no missing values in this dataset.

3.4 Overview of Dataset

Table 1: Summary Numeric Variables

	N	Mean	SD	Min	Q1	Median	Q3	Max
Age	1470	36.92	9.14	18	30	36.0	43	60
DailyRate	1470	802.49	403.51	102	465	802.0	1157	1499
DistanceFromHome	1470	9.19	8.11	1	2	7.0	14	29
EmployeeNumber	1470	1024.87	602.02	1	491	1020.5	1556	2068
HourlyRate	1470	65.89	20.33	30	48	66.0	84	100
MonthlyIncome	1470	6502.93	4707.96	1009	2911	4919.0	8380	19999
MonthlyRate	1470	14313.10	7117.79	2094	8045	14235.5	20462	26999
NumCompaniesWorked	1470	2.69	2.50	0	1	2.0	4	9
PercentSalaryHike	1470	15.21	3.66	11	12	14.0	18	25
TotalWorkingYears	1470	11.28	7.78	0	6	10.0	15	40
TrainingTimesLastYear	1470	2.80	1.29	0	2	3.0	3	6
YearsAtCompany	1470	7.01	6.13	0	3	5.0	9	40
YearsInCurrentRole	1470	4.23	3.62	0	2	3.0	7	18
YearsSinceLastPromotion	1470	2.19	3.22	0	0	1.0	3	15
YearsWithCurrManager	1470	4.12	3.57	0	2	3.0	7	17

Table 2: Summary Factor Variables

	Level	N	%		Level	N	%
Attrition	No	1233	83.9		5	69	4.7
	Yes	237	16.1	JobRole	Healthcare Representative	131	8.9
BusinessTravel	None	1043	71.0		Human Resources	52	3.5
	Rarely	150	10.2		Laboratory Technician	259	17.6
	Frequently	277	18.8		Manager	102	6.9
Department	HR	63	4.3		Manufacturing Director	145	9.9
	R&D	961	65.4		Research Director	80	5.4
	Sales	446	30.3		Research Scientist	292	19.9
Education	Below College	170	11.6		Sales Executive	326	22.2
	College	282	19.2		Sales Representative	83	5.6
	Bachelor	572	38.9	JobSatisfaction	Low	289	19.7
	Master	398	27.1		Medium	280	19.0
	Doctor	48	3.3		High	442	30.1
EducationField	Human Resources	27	1.8		Very High	459	31.2
	Life Sciences	606	41.2	MaritalStatus	Divorced	327	22.2
	Marketing	159	10.8		Married	673	45.8
	Medical	464	31.6		Single	470	32.0
	Other	82	5.6	OverTime	No	1054	71.7
	Technical Degree	132	9.0		Yes	416	28.3
EnvironmentSatisfaction	Low	284	19.3	PerformanceRating	Excellent	1244	84.6
	Medium	287	19.5		Outstanding	226	15.4
	High	453	30.8	RelationshipSatisfaction	Low	276	18.8
	Very High	446	30.3		Medium	303	20.6
Gender	Female	588	40.0		High	459	31.2
	Male	882	60.0		Very High	432	29.4
JobInvolvement	Low	83	5.6	StockOptionLevel	0	631	42.9
	Medium	375	25.5		1	596	40.5
	High	868	59.0		2	158	10.7
	Very High	144	9.8		3	85	5.8
JobLevel	1	543	36.9	WorkLifeBalance	Bad	80	5.4
	2	534	36.3		Good	344	23.4
	3	218	14.8		Better	893	60.7
	4	106	7.2		Best	153	10.4

3.5 Splitting Dataset (Stratification)

There is significant class imbalance in the variable Attrition (84 / 16) as can be seen in Table 2, which makes sense since most employees want to work at the company. However it is important to stratify the train and test split so that we receive a more realist estimate on how our model is going to perform.

```
set.seed(111)
emp_attrition_split <- initial_split(emp_attrition, prop = 0.80, strata = "Attrition")
emp_attrition_train <- training(emp_attrition_split) # i.e. 986 / 189 = 5.22
emp_attrition_test  <- testing(emp_attrition_split) # i.e. 248 / 48 = 5.17
```

In this analysis we are splitting the data set with 80% training and 20% test / validation. As can be observed above we have an almost equal ratio of yes to no in the training as well as the testing data set due to stratification.

4 Exploration

4.1 Jobs

There seems to be different Job positions for different departments.

Human Resources Department

- Human Resources Managers
- Human Resources

Research & Development Department

- Research Scientist
- Laboratory Technician

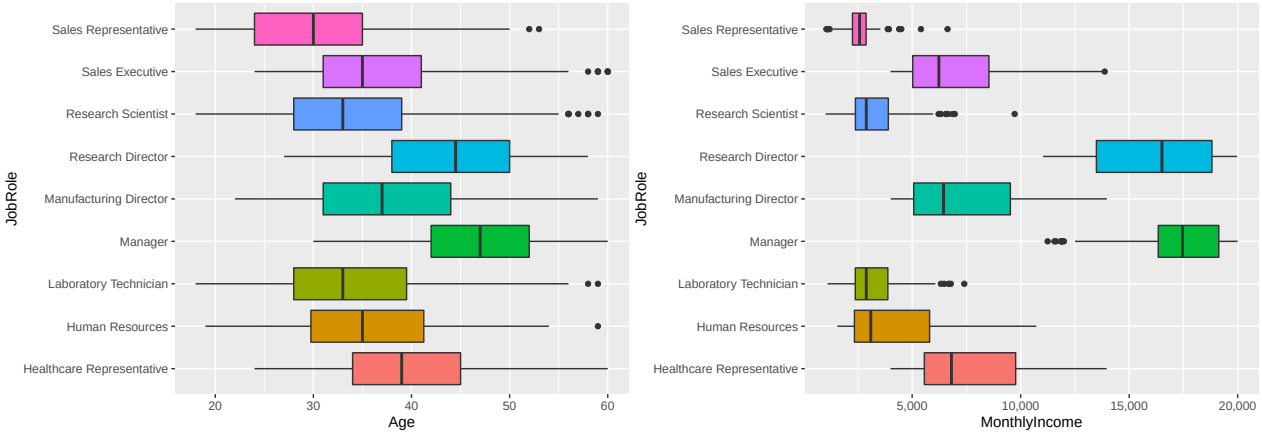
- Manufacturing Director
- Healthcare Representative
- Research Director
- Research Manager

Sales Department

- Sales Executive
- Sales Representative
- Sales Manager

It is clear that the company has quite a hierarchical structure and possibly organigram as different job levels have different monthly income which increase in each job role depending on the job level reached. Furthermore, the job level one segment starts around 2'500 which is most likely for entry positions and reaches up to around 20 thousand monthly for executive positions.

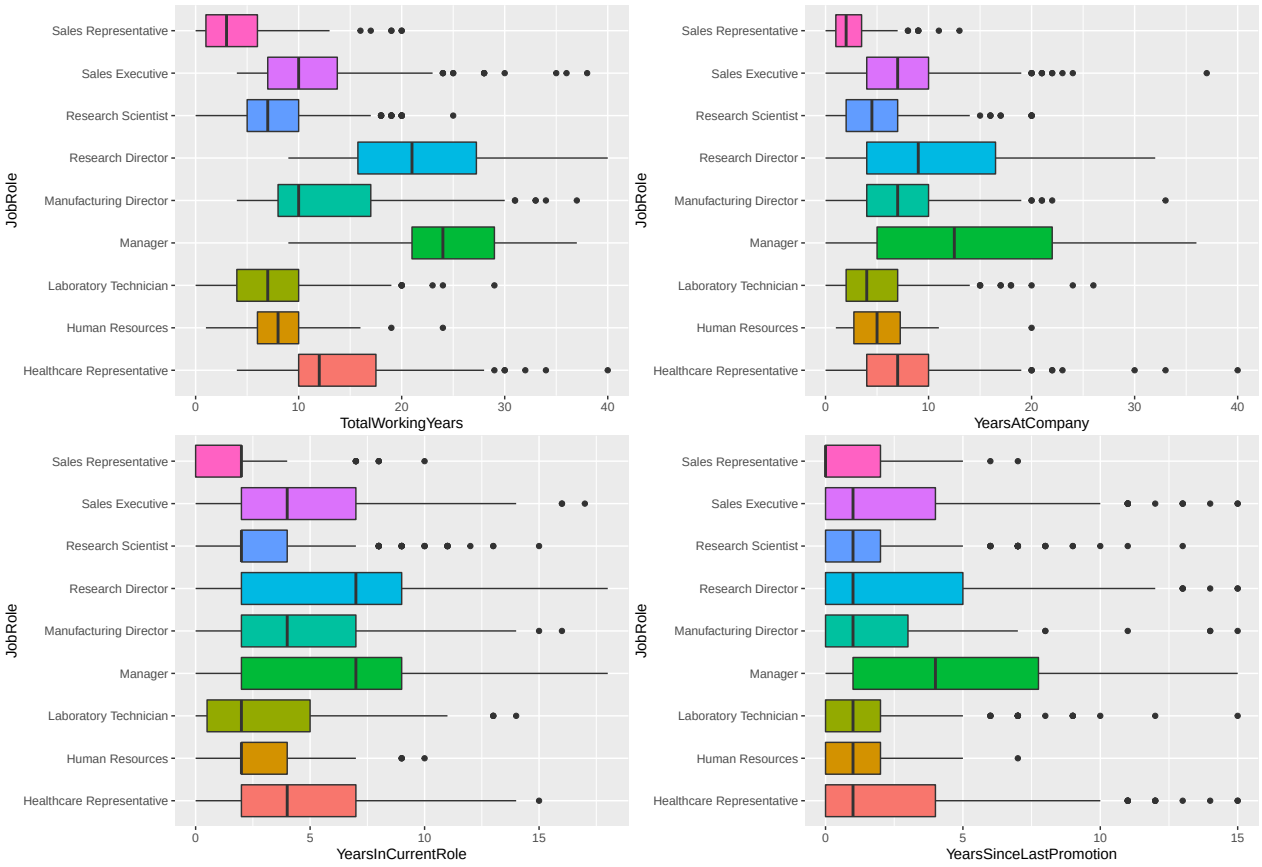
4.2 Age



Based on the above box plots one can take a few things away.

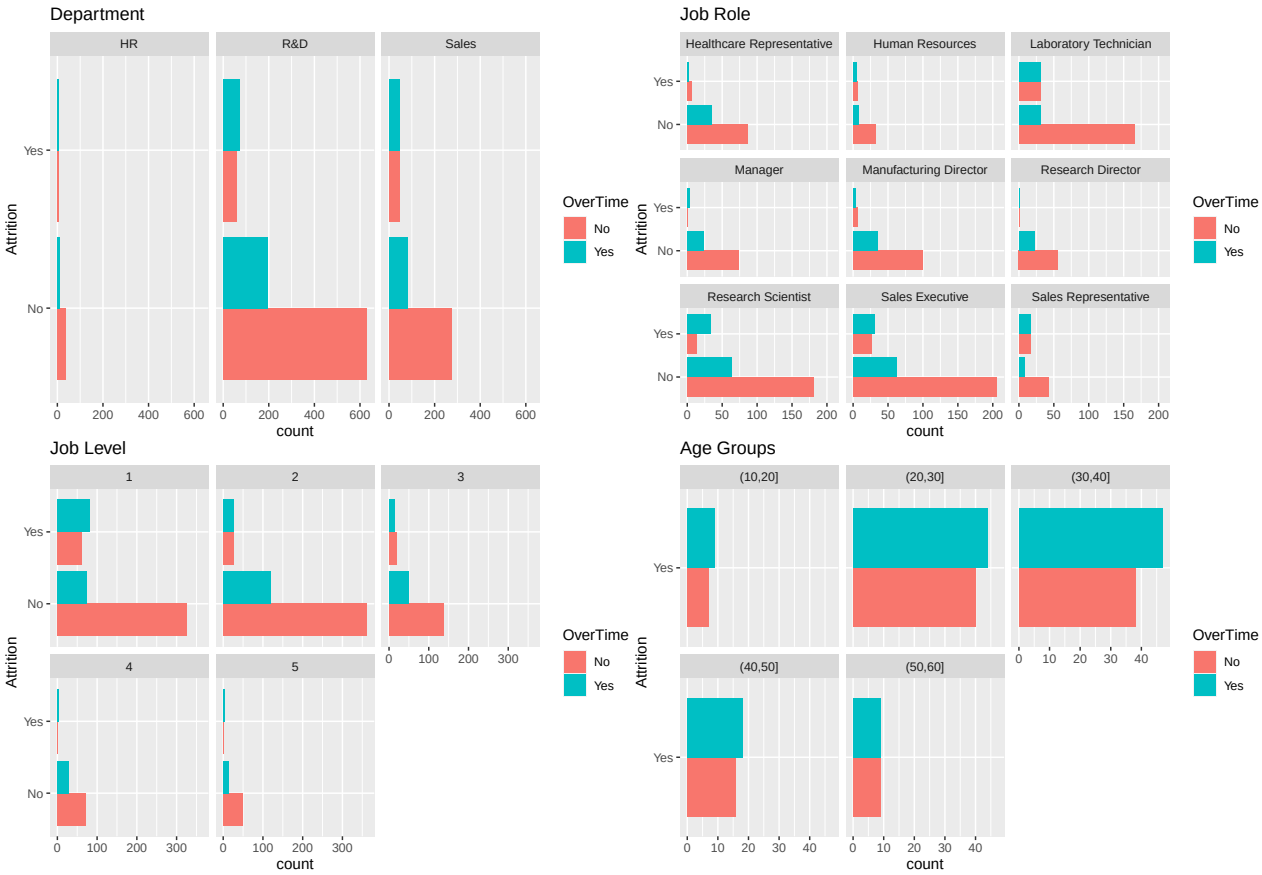
- Manager's and Research Director's have the oldest median age
- They are making the highest median monthly income as well
- On average sales representatives are the youngest in this dataset and are making the lowest monthly income, this could possibly also be external employee's

4.3 Working Experience



- Manager's and Research Director's have the highest total working years and years at the current company, which again makes us think of a very hierarchical company
- Sales Representatives have the lowest total working years and years at the current company, which leads one to think they either quickly being promoted or don't like the job
- Years since last promotion is more or less the sames for all the job roles other than managers, which is good sign that people are not leaving the company just because of lacking promotions

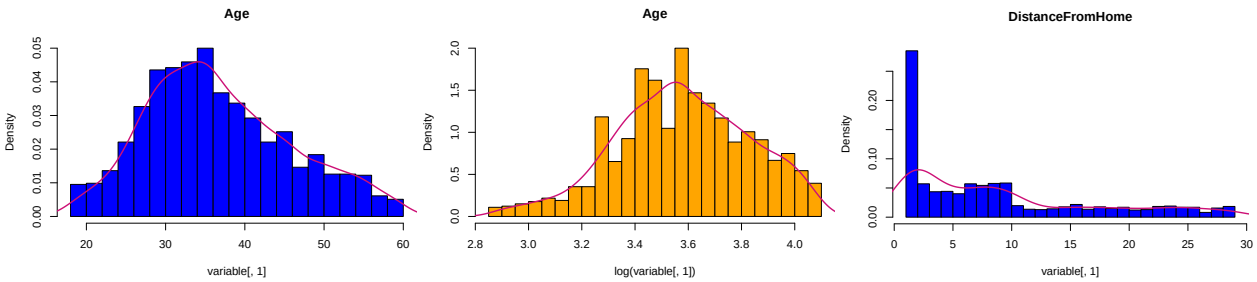
4.4 Over Time and Employee Attrition

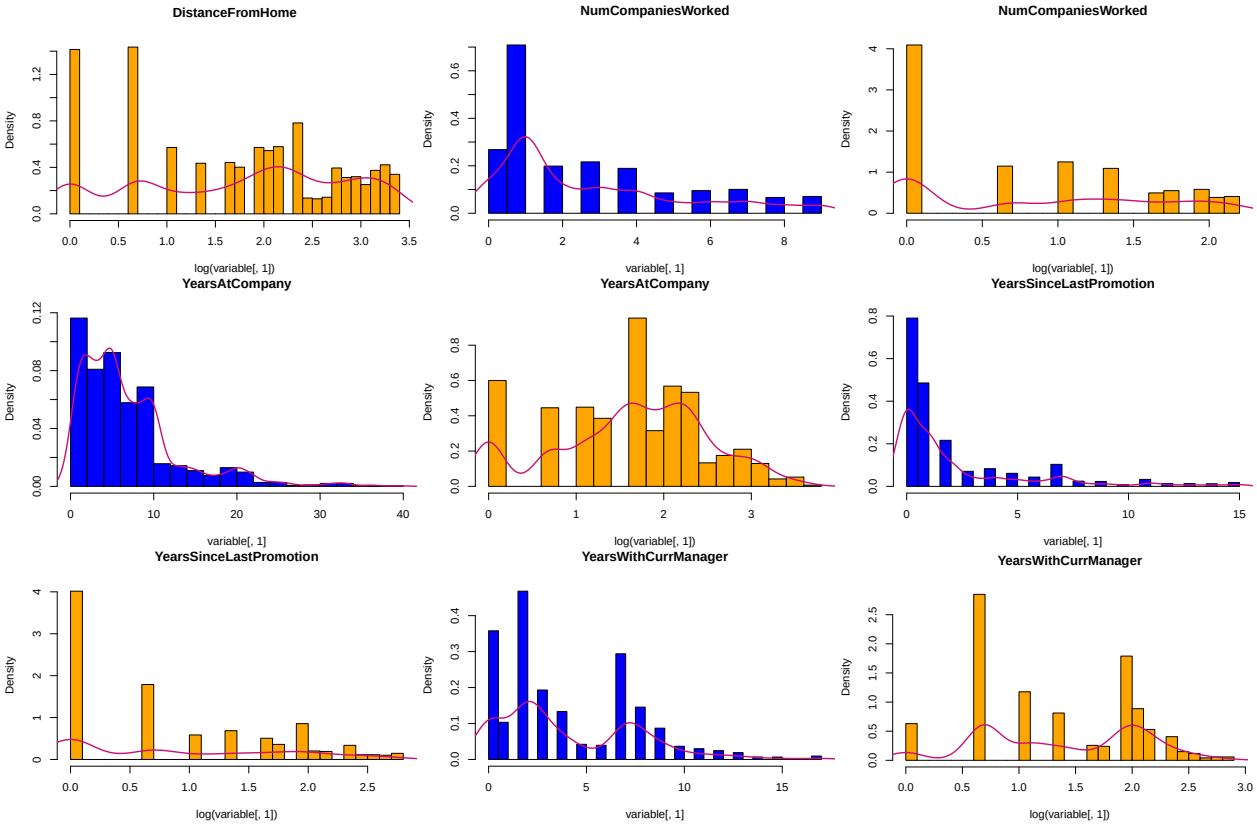


Overall as can be seen in the departments people with overtime have been leaving the company. Moreover it seems R&D and Sales department are most affected by this. When breaking it down even further to job roles one quickly see's that research scientist are most likely to quite due to overtime. In terms of positions entry job level positions are also highly sensitive to overtime. There also seems to be a trend for Senior and Executives, which could hint at a burnout rate. When considering age groups as well people seem to be leaving the company with overtime when they are in their thirties because they are starting a family or have already quite a bit of work experience and quickly may find an other job.

4.5 Histograms of variables

In this section the distributions of the numerical variables are analyzed visually, also in order to evaluate if a variable needs to be log transformed. Therefore, density histograms are displayed for each variable with log transformation(orange) and without transformation(blue).





Findings Histograms:

- All variables, expect the variable "Years with current Manager", show some sort of a right skeewed distribution and might be log transformed.
- The variable "Years with current Manager" seems to display a bimodal distribution with two peaks, with the second peak showing right skeewedness. For a linear model, the values of the variable could be split around the value 5 to analyse each peak separately, however, this might be most applicable if the residuals of the linear model also display a bimodal distribution indicating a strong influence of the variable.
- The variables "distance from home", "number of companies worked" and "years since last promotion" do not display a normal distribution after the log transformation. This will be further considered when building a linear model.

5 Linear Regression

As it does not make sense to fit a Linear Model to a categorical variable such as Attrition (Levels: Yes/No), the Linear Models and also the General Additive Models will be fitted using YearsAtCompany as the dependent variable. We assume that the longer employees work for a company, the less attrition would actually happen. Therefore, the insight of this chapter might help to fit more complex model to the variable Attrition in the following chapters, but also indicate what leads employees to work longer for an employer.

5.1 Simple Linear Model

Before performing a Linear Regression Model, the reference level for the factor variables are set to the level with the greatest n-count. Due to the length of the output, only the factor levels of the first two factor variables are displayed here.

```
## variable: Attrition, factor_levels: c(No = 986, Yes = 189)
```

```
## variable: BusinessTravel, factor_levels: c(None = 832, Rarely = 130, Frequently = 213)
```

The first Linear Model is performed including all variables without additional log transformation. As the output is quite long it is omitted here.

```
# Linear model without log transformation
lm.YearsAtCompany.0 <- lm(formula=YearsAtCompany ~ ., data = emp_attrition_train_lm)
lm.summary.0 <- summary(lm.YearsAtCompany.0)
lm.summary.0
```

Findings LM Model .0:

- According to the Adjusted R-squared value the model explains 68.48% of the variance of the dependent variable.
- Variables with a significant influence on the dependet variable YearsAtCompany, according to a significance level of 0.05:
  - Age, NumCompaniesWorked, YearsSinceLastPromotion, YearsWithCurrManager
- The specific influence of each variable will be interpreted at a later stage for a better fitting model.

5.2 Linear Model with Log Transformation

In the next step the independent variables (Age, YearsAtCompany, YearsWithCurrentManager) which displayed right skeewedness are log transformed for a further Linear Model. As the log transformation did create infinity values from log(0) those values are changed back to 0. No NaN values are created because there are no negative values in the transformed variables. Created NaN values due to log transformation would most likely need to be deleted from a dataset if they can not be replaced by a reasonable value. Furthermore, the reference levels have to be set again for the newly created _log dataset. The output of the log transformed model is omitted here as it is rather long.

```
# LM All variables, incl. log transformation
lm.YearsAtCompany.1 <- lm(formula = YearsAtCompany ~ ., data = emp_attrition_log)
lm.summary.1 <- summary(lm.YearsAtCompany.1)
lm.summary.1
```

Findings LM Models .1 including log-transformation:

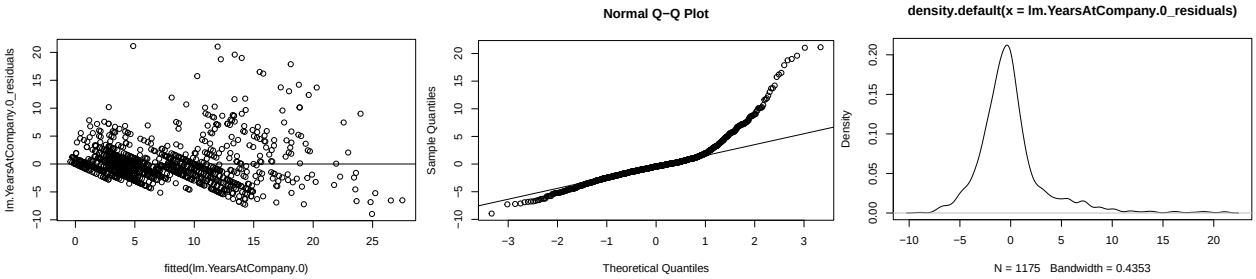
- Variables with a significant influence on the dependet variable YearsAtCompany, according a significance level of 0.05:
  - Attrition, Age, NumCompaniesWorked, YearsSinceLastPromotion, YearsWithCurrManager.
  - Compared to the previous model the variable Attrition became additionally significant.

- According to the Adjusted R-squared value, the model explains 71,98% of the variance of the dependent variable displaying an increase of 3.50% compared to the previous model without log transformation.
- Interestingly, the categorical variables, EnvironmentSatisfaction, WorkLifeBalanceGood, display almost significant p-values for at least one factor level and will be kept for further modelling too.
- The variable YearsWithCurrentManager was log transformed due to the formely detected right skeewedness of the second peak, which can be reduced by this procedure. Furthermore, it increased also the adjusted R-Squared value of the model.
- The variables DistanceFromHome, NumberOfCompaniesWorked and YearsSinceLastPromotion are not log transformed because this decreased the obtained Adjusted R-squared value in a trial.

5.3 Residual Analysis LM Models

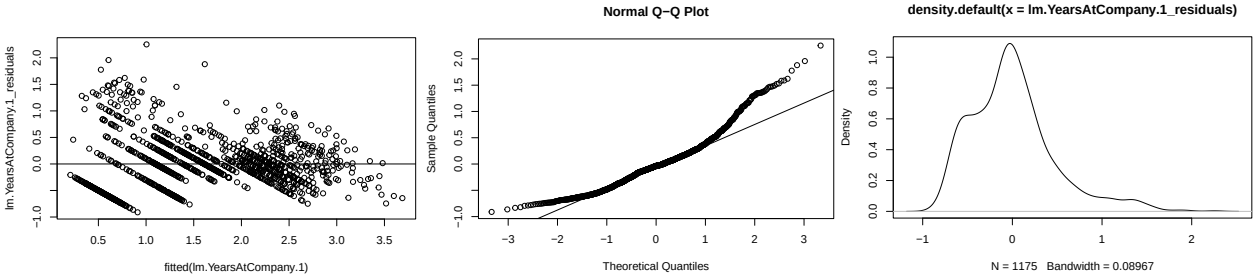
Without log transformation:

- *Resiudals vs. fitted plot*: The spread of residuals seems to be higher for higher fitted values. Therefore, the normal distributions should be further questioned.
- *QQ-Plot*: The strayed residuals around higher and lower values indicate that they might not be normally distributed.
- *Density plot*: As the residuals are mainly bell shaped, but also displaying right skeewedness, the normal distribution will still be assumed. Furthermore, the formerly mentioned binomial distribution of the variable YearsWithCurrentManager seems not to influence the distribution of the residuals heavily. Therefore, the two peaks of the variable will not be analyzed separately nor will the data of the variable be split between the peaks. A split of the dataset would of course also have further implications which can be avoided in this analysis.



With log transformation:

- *Residuals vs. fitted plot*: The spread of residuals for higher fitted values was reduced compared to the model without log transformation. However, the opposite can be observed for lower fitted values.
- *QQ-Plot*: In this plot the residuals seem better distributed but still strayed for higher values. For lower values the assumption of normal distribution seems less appropriate now.
- *Density plot*: The bell shape is still somehow given, the length of the tails got reduced relatively to the scale.



5.4 Final Linear Model and Verification

Final Linear Model:

For the Final Linear Model only the formerly significant variables are kept, including categorical variables that did display at least one almost significant level.

```
##
## Call:
## lm(formula = YearsAtCompany ~ Attrition + Age + EnvironmentSatisfaction +
##     NumCompaniesWorked + WorkLifeBalance + YearsSinceLastPromotion +
##     YearsWithCurrManager, data = emp_attrition_log)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.92065 -0.33361 -0.04173  0.20562  2.24725
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -0.621477   0.218485  -2.844 0.004526 **
## AttritionYes                    -0.147561   0.039707  -3.716 0.000212 ***
## Age                              0.372987   0.062859   5.934  3.9e-09 ***
## EnvironmentSatisfactionLow      -0.072942   0.040987  -1.780 0.075399 .
## EnvironmentSatisfactionMedium   -0.038294   0.040276  -0.951 0.341904
## EnvironmentSatisfactionVery High -0.028366   0.035474  -0.800 0.424085
## NumCompaniesWorked              -0.023672   0.006076  -3.896 0.000103 ***
## WorkLifeBalanceBad               0.007476   0.061902   0.121 0.903897
## WorkLifeBalanceGood              0.059313   0.034003   1.744 0.081365 .
## WorkLifeBalanceBest              0.086010   0.047678   1.804 0.071492 .
## YearsSinceLastPromotion          0.053543   0.004984  10.743  < 2e-16 ***
## YearsWithCurrManager             0.746918   0.019568  38.171  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.477 on 1163 degrees of freedom
## Multiple R-squared:  0.7234, Adjusted R-squared:  0.7208
## F-statistic: 276.5 on 11 and 1163 DF,  p-value: < 2.2e-16
```

Interpretation:

- Adjusted R-squared value: The model is explaining 72.08% of the variance of the dependent variable YearsAtCompany. Compared to the former model including all variables this accounts for a small increase of 0.01% by omitting four variables.
- Intercept: YearsAtCompany if all other variables were 0 or at the reference level. Given variables such as Age, it is pointless to interpret the intercept for this model.

Interpretation of log transformed predictors on log transformed dependent variable:

- Age: The dependent variable, YearsatCompany, increases by 0.37% years if the Age of an employee increases by 1%. If for example a 40 years old and a 44 years old employee are compared (+10%), the older employee is expected to have worked 3.70% longer at the company.
- YearsWithCurrManager: The dependent variable, YearsAtCompany, increases by 0.75% if the independent variable increases by 1%. It is not surprising that people seem to prefer consictency in their direct management.

Interpretation of numerical predictors on log transformed dependent variable:

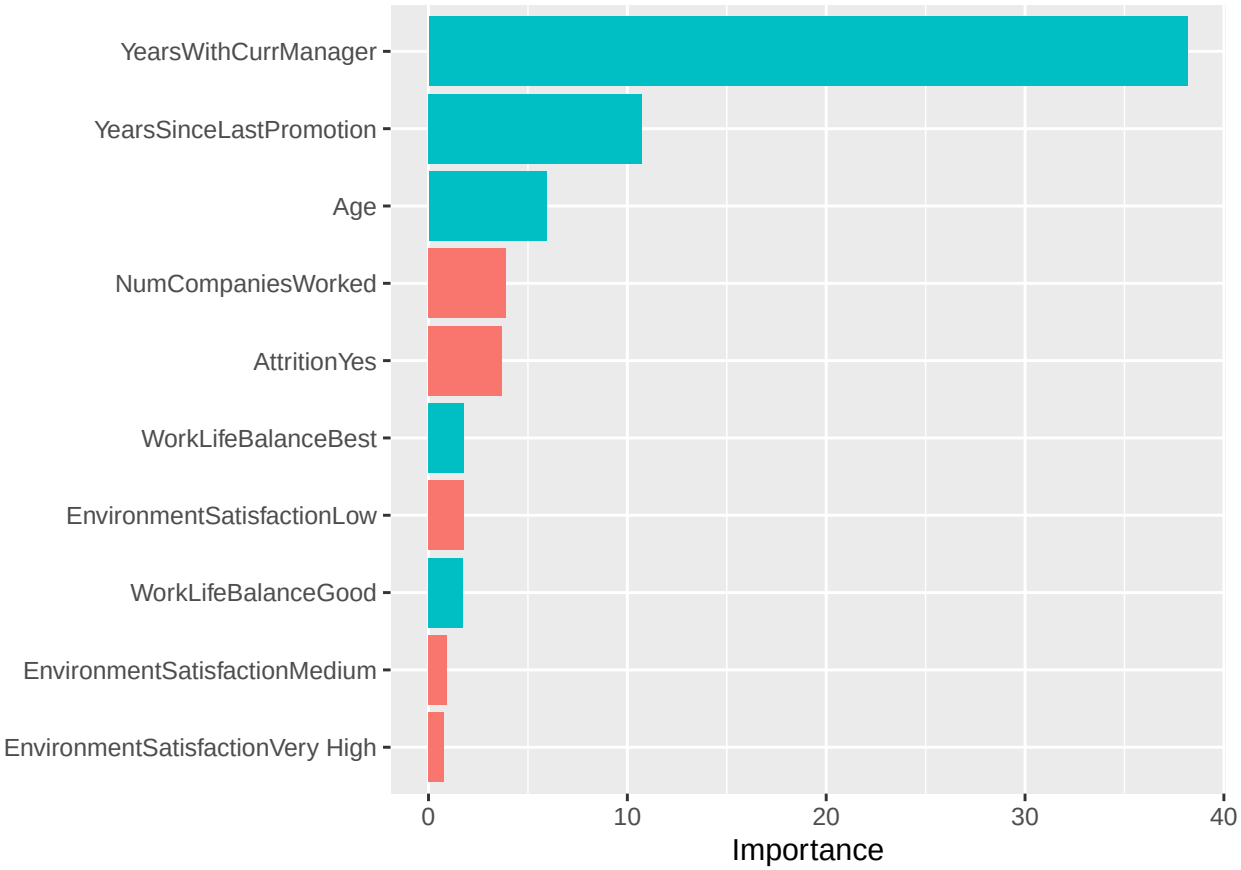
- NumCompaniesWorked: The dependent variable, YearsAtCompany, decreases by -2.34% if the independent variable increases by 1 company. People who changed their employer often in the past might continue to do so.
- YearsSinceLastPromotion: The dependent variable, YearsAtCompany, increases by 5.50% if the independent variable increases by 1 year. People who might be waiting/expecting a promotion are likely to work longer for a company.

Interpretation of categorical predictors on log transformed dependent variable, relevant levels only:

- EnvironmentSatisfaction: The dependent variable, YearsAtCompany, decreases by 7.03% if the EnvironmentSatisfaction is "low", compared to the reference level "high". People with low EnvironmentSatisfaction are likely to leave the company sooner compared to people with high EnvironmentSatisfaction.
- WorkLifeBalance: The dependent variable, YearsAtCompany, increases by 6.11% respectively by 8.98% if the balance is "good" or even "best", compared to the reference level "better". The better the WorkLifeBalance of employees, the longer they seem to keep working for the company.
- Attrition: The dependent variable, YearsAtCompany, decreases by 13.72% if an employee left the company. The decrease is not surprising but does not provide much additional insights with the available information.

Variable importance plot:

In the following plot the variables are sorted by importance, variables that influence the dependent variable YearsAtCompany positively are colored cyan, variables that influence it negatively are colored red.



Verification:

As a first verification step a strict model is fitted, only including the variables below the significance level 0.05.

```
##
## Call:
## lm(formula = YearsAtCompany ~ Attrition + Age + NumCompaniesWorked +
##     YearsSinceLastPromotion + YearsWithCurrManager, data = emp_attrition_log)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8483 -0.3290 -0.0458  0.2156  2.3131
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -0.611730   0.217239  -2.816 0.004945 **
## AttritionYes            -0.153689   0.039140  -3.927 9.12e-05 ***
## Age                      0.368238   0.062818   5.862 5.94e-09 ***
## NumCompaniesWorked      -0.023250   0.006066  -3.833 0.000133 ***
## YearsSinceLastPromotion  0.054087   0.004980  10.862  < 2e-16 ***
## YearsWithCurrManager     0.745900   0.019567  38.120  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4775 on 1169 degrees of freedom
## Multiple R-squared:  0.7213, Adjusted R-squared:  0.7201
## F-statistic: 605.2 on 5 and 1169 DF,  p-value: < 2.2e-16
```

Interpretation:

- Only slight decrease of adjusted R-squared compared to the previous model by excluding 3 variables.
- However, when looking at the "Residual Sums of Squares" (RSS) in the *Analysis of Variance Table* below it becomes apparent that the previous model did explain slightly more of the variance of the dependent variable YearsAtCompany. Therefore, the more complex final model will be investigated further in the next section using a General Additive Models.

```
## Analysis of Variance Table
##
## Model 1: YearsAtCompany ~ Attrition + Age + EnvironmentSatisfaction +
##     NumCompaniesWorked + WorkLifeBalance + YearsSinceLastPromotion +
##     YearsWithCurrManager
## Model 2: YearsAtCompany ~ Attrition + Age + NumCompaniesWorked + YearsSinceLastPromotion +
##     YearsWithCurrManager
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   1163 264.64
## 2   1169 266.59 -6   -1.9552 1.4321 0.1989
```

Verification of the final model with the test data set:

To test the final model with the test data, the test data set is equivalently transformed as the training data set. This includes the log transformation of 3 variables, including the dependent variable, setting infinity values to 0 and setting the reference level to the level with the greatest n-count.

```
##
## Call:
## lm(formula = YearsAtCompany ~ Attrition + Age + EnvironmentSatisfaction +
```

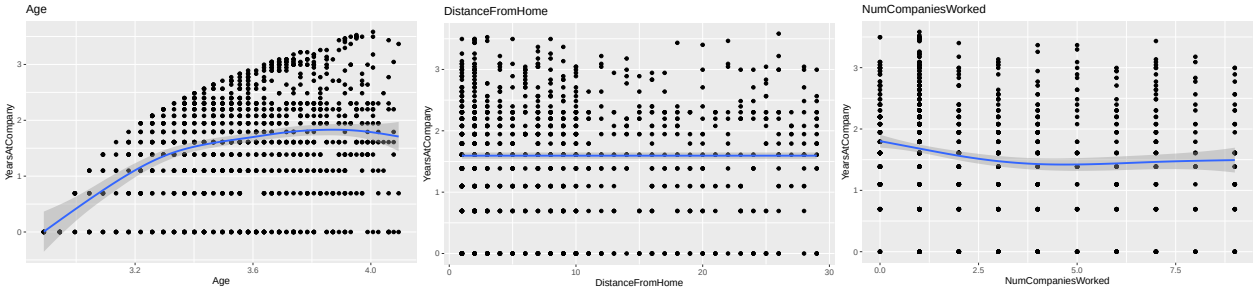
```
##     NumCompaniesWorked + WorkLifeBalance + YearsSinceLastPromotion +
##     YearsWithCurrManager, data = emp_attrition_test_lm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.89739 -0.31219 -0.08144  0.19436  1.73297
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      -0.51148    0.42497  -1.204  0.22976
## AttritionYes                     -0.14796    0.07989  -1.852  0.06505 .
## Age                               0.34046    0.12164   2.799  0.00548 **
## EnvironmentSatisfactionLow        0.10371    0.08052   1.288  0.19879
## EnvironmentSatisfactionMedium     0.03688    0.08266   0.446  0.65584
## EnvironmentSatisfactionVery High -0.05880    0.07350  -0.800  0.42439
## NumCompaniesWorked               -0.03398    0.01160  -2.930  0.00367 **
## WorkLifeBalanceBad               -0.20469    0.13559  -1.510  0.13226
## WorkLifeBalanceGood               0.03060    0.06820   0.449  0.65405
## WorkLifeBalanceBest              -0.01440    0.08891  -0.162  0.87142
## YearsSinceLastPromotion           0.05833    0.01019   5.727 2.61e-08 ***
## YearsWithCurrManager              0.77200    0.03978  19.405  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4823 on 283 degrees of freedom
## Multiple R-squared:  0.741,  Adjusted R-squared:  0.731
## F-statistic: 73.61 on 11 and 283 DF,  p-value: < 2.2e-16
```

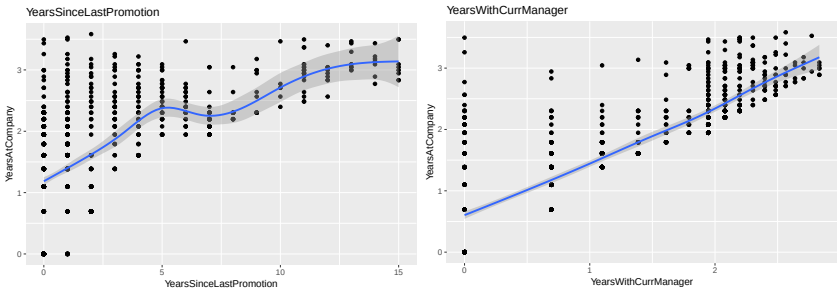
Interpretation:

- The model reaches a slighty higher Adjusted R-squared value of 73.10% applied on the test data set.
- However, the variable Attrition ist just barely not significant anymore when considering the 0.05 significance level and also the other categorical variables loose in significance even more strongly.
- Overall, the model seems to be still applicable to the test set, however, as the test data set is significantly smaller than the training data set, some effects seem to fade.

6 General Additive Model

Generalised Additive Models (GAMs) are an adaptation that allows to model non-linear data while maintaining explainability. In a first step the relationship between the dependent variable YearsAtCompany with the independent variables are analyzed visually and interpreted to verify whether a smooth term, e.g. for a quadratic or cubic relationship, would need to be included in a model. As it is an extension of the linear model, the variables Age, YearsAtCompany and YearsWithCurrentManager are kept as log transformed variables.





Interpretation of relationship of YearAtCompany ~ Variable:

- Age: The relationship might be quadratic.
- DistancefromHome: The relationship seems to be constant. It therefore does not surprise that this variable was excluded in the previous linear model. It will be excluded from the General Additive Model as well.
- NumCompaniesWorked: The relationship might be quadratic.
- YearSinceLastPromotion: The relationship seems to be rather complex than linear.
- YearsWithCurrManager: The relationship seems to be more or less linear but with a few imperfections.

General Additive model with smooth terms:

When the General Additive model is fitted to the same data set as the final model, assuming a "Gaussian" distribution, it returns the same output as for the final linear model. Therefore, we include smooth terms for the numerical variables in the following model. When testing including smooth terms variable by variable, it did increase the Adjusted R-squared value each time. By design, the model does choose the appropriate degree of complexity for the smooth terms itself as it is unknown to the user in most cases.

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## YearsAtCompany ~ Attrition + s(Age) + EnvironmentSatisfaction +
##     s(NumCompaniesWorked) + WorkLifeBalance + s(YearsSinceLastPromotion) +
##     s(YearsWithCurrManager)
##
## Parametric coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      1.610464   0.026661  60.406  < 2e-16 ***
## AttritionYes                    -0.128273   0.038761  -3.309 0.000964 ***
## EnvironmentSatisfactionLow      -0.047619   0.039849  -1.195 0.232344
## EnvironmentSatisfactionMedium   -0.030994   0.039019  -0.794 0.427175
## EnvironmentSatisfactionVery High -0.013050   0.034400  -0.379 0.704487
## WorkLifeBalanceBad               0.007009   0.060120   0.117 0.907215
## WorkLifeBalanceGood              0.063681   0.032985   1.931 0.053779 .
## WorkLifeBalanceBest              0.086620   0.046493   1.863 0.062711 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                            edf Ref.df       F  p-value
## s(Age)                   4.914  6.003  11.370  < 2e-16 ***
## s(NumCompaniesWorked)    2.660  3.331   8.646 6.98e-06 ***
```

```
## s(YearsSinceLastPromotion) 7.722  8.494  20.340  < 2e-16 ***
## s(YearsWithCurrManager)    7.092  7.956 171.713  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.742   Deviance explained = 74.8%
## GCV =  0.216  Scale est. = 0.21041   n = 1175
```

Interpretation:

- The model reaches an Adjusted R squared value of 74.20%, which is slightly higher as the final linear model.
- Variables with a significant effect on the dependent variable according to a significance level of 0.05:
  - Attrition, Age, NumCompaniesWorked, YearsSinceLastPromotion and YearsWithCurrentManager just as in the final linear model.
- For the numerical predictors Age, NumCompaniesWorked, YearsSinceLastPromotion and YearsWithCurrentManager, there is a strong evidence, that the variables have a non-linear effect on the dependent variable, according to the estimated degrees of freemdom (edf). A edf of 1 would indicate a linear relationship.
- As there seem to be strong non-linear effects, the influence of the numerical variables can not be interpreted as simple as for the linear model and will not be regarded here.

In the next step this General Additive Model is verified by using the test-dataset, for which the log transformation and setting of reference levels was perfomed at an earlier stage.

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## YearsAtCompany ~ Attrition + s(Age) + EnvironmentSatisfaction +
##     s(NumCompaniesWorked) + WorkLifeBalance + s(YearsSinceLastPromotion) +
##     s(YearsWithCurrManager)
##
## Parametric coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      1.561636   0.058267  26.802   <2e-16 ***
## AttritionYes                    -0.121745   0.079552  -1.530    0.127
## EnvironmentSatisfactionLow       0.091574   0.079680   1.149    0.251
## EnvironmentSatisfactionMedium    0.013933   0.082290   0.169    0.866
## EnvironmentSatisfactionVery High -0.072371   0.072790  -0.994    0.321
## WorkLifeBalanceBad               -0.203518   0.133942  -1.519    0.130
## WorkLifeBalanceGood               0.031823   0.067431   0.472    0.637
## WorkLifeBalanceBest               0.001128   0.088147   0.013    0.990
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                              edf Ref.df       F p-value
## s(Age)                     1.921  2.423   5.003 0.00486 **
## s(NumCompaniesWorked)      1.465  1.792   6.640 0.00426 **
## s(YearsSinceLastPromotion) 1.000  1.000  37.558 < 2e-16 ***
## s(YearsWithCurrManager)    2.016  2.498 150.199 < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.738   Deviance explained =   75%
## GCV = 0.23799  Scale est. = 0.22638   n = 295
```

Interpretation:

- The Adjusted R-squared value is close to the equivalent value of the GAM performed on the training data set.
- However, the variable Attrition is not significant anymore if a significance level of 0.05 is applied, along with the categorical variables EnvironmentSatisfaction and WorkLifeBalance loosing in significance too. The same effects were noticed when testing the final linear model on the smaller test data set.
- Interestingly, the smoothed numerical variables are still significant but display quite different edf values. For the variable YearsSinceLastPromotion a simple linear relationship would have been adequate for the test model, altough, a smooth term seemed appropriate in the trial for the previous General Additive Model.

Evaluation GAM vs. Final Linear Model:

Overall, the general additive model performs slighty better on the training dataset and also when applied to the test dataset, if measured by the returned Adjusted R-squared values. Similar effects are displayed for the fading of influence of the categorical variables Attrition, EnvironmentSatisfaction and WorkLifeBalance on the dependent variable YearsAtCompany for both approaches, when fitted to the smaller test dataset. The smoothed terms in the General Additive model seem to fit the numerical variables slightly better, however, taking into account that the effects of the variables can be interpeted quite straightforward in the linear model, the final linear model seems to fit the attrition data sets sufficiently.

7 Generalised Linear Models

7.1 Poisson Regression

```
glm.companies.worked.1 <- glm(NumCompaniesWorked ~ Age, family = "quasipoisson", data = emp_attrition)
glm.companies.worked.2 <- glm(NumCompaniesWorked ~ Age + Attrition + OverTime + TotalWorkingYears + Yea
# summary(glm.companies.worked.2)
print(exp(coef(glm.companies.worked.2)["Age"]), digits = 5)
```

```
##    Age
## 1.0196
```

```
print(exp(coef(glm.companies.worked.2)["AttritionYes"]), digits = 5)
```

```
## AttritionYes
##       1.2827
```

```
print(exp(coef(glm.companies.worked.2)["OverTimeYes"]), digits = 5)
```

```
## OverTimeYes
##     0.88247
```

```
print(exp(coef(glm.companies.worked.2)["TotalWorkingYears"]), digits = 5)
```

```
## TotalWorkingYears
##            1.0312
```

```
print(exp(coef(glm.companies.worked.2)["YearsAtCompany"]), digits = 5)
```

```
## YearsAtCompany
##        0.95287
```

Fitting a poisson distribution on this data set is done purely for the academic purposes as since number of companies worked for won't fit poisson distribution therefore the quasipoisson is used. Nevertheless the following interpretation can be made.

For a given employee,

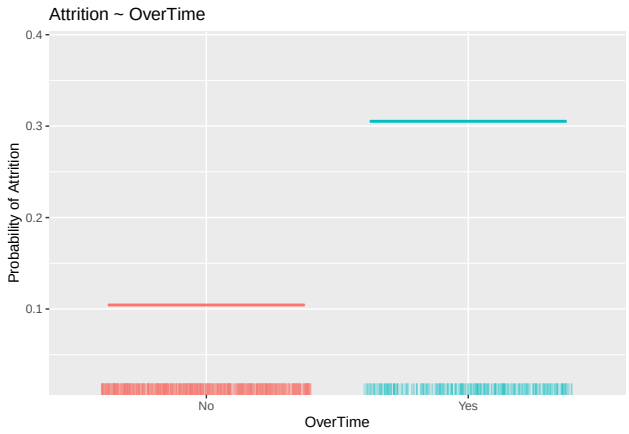
- increasing the age by one year, would result in about 2% more number of companies worked for
- Employee who are leaving the IBM company, on average, have 28% number of companies worked for than people who are not leaving the company
- Employee with overtime, on average, have 88% more number of comapnies worked for than peiple who are not doing overtime
- for each year a employee works about 3% more number of companies worked for increases
- for each year a employee works for IBM the number of companies worked for decreases by 5%

7.2 Simple Logistic Regression

```
attrition_model1 <- glm(Attrition ~ OverTime, family = "binomial", data = emp_attrition)
summary(attrition_model1)$coefficients
```

```
##              Estimate Std. Error    z value      Pr(>|z|)
## (Intercept) -2.149646  0.1007431 -21.337888 5.052603e-101
## OverTimeYes  1.327406  0.1465721   9.056333  1.349094e-19
```

Not unexpectedly overtime seems to play a relevant role. Indeed, its p-value is highly significant.

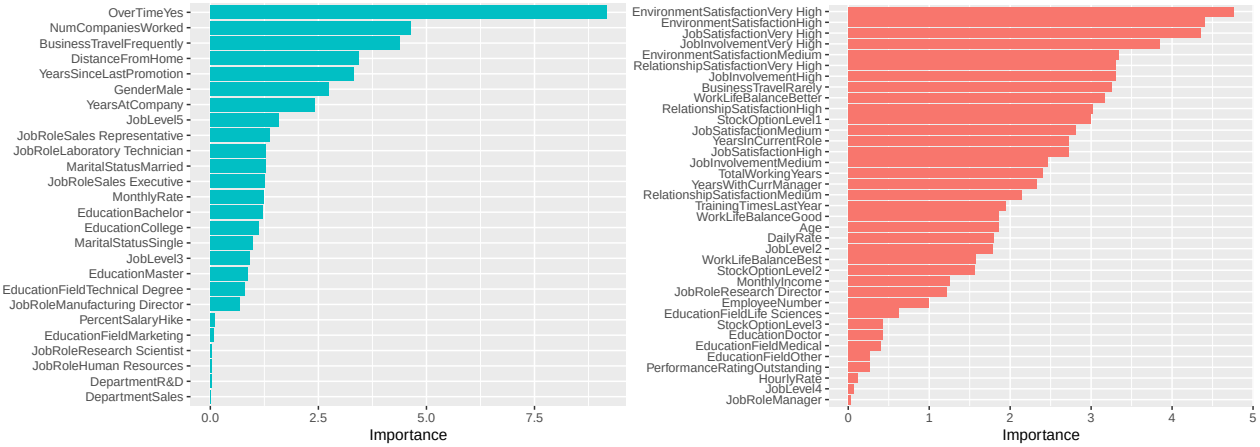



```
exp(coef(attrition_model1)["OverTimeYes"])
```

```
## OverTimeYes
##    3.771249
```

The odds of someone leaving the company with overtime are about ~3.8 times higher than the odds for no overtime in this simple model.

7.3 Multiple logistic regression

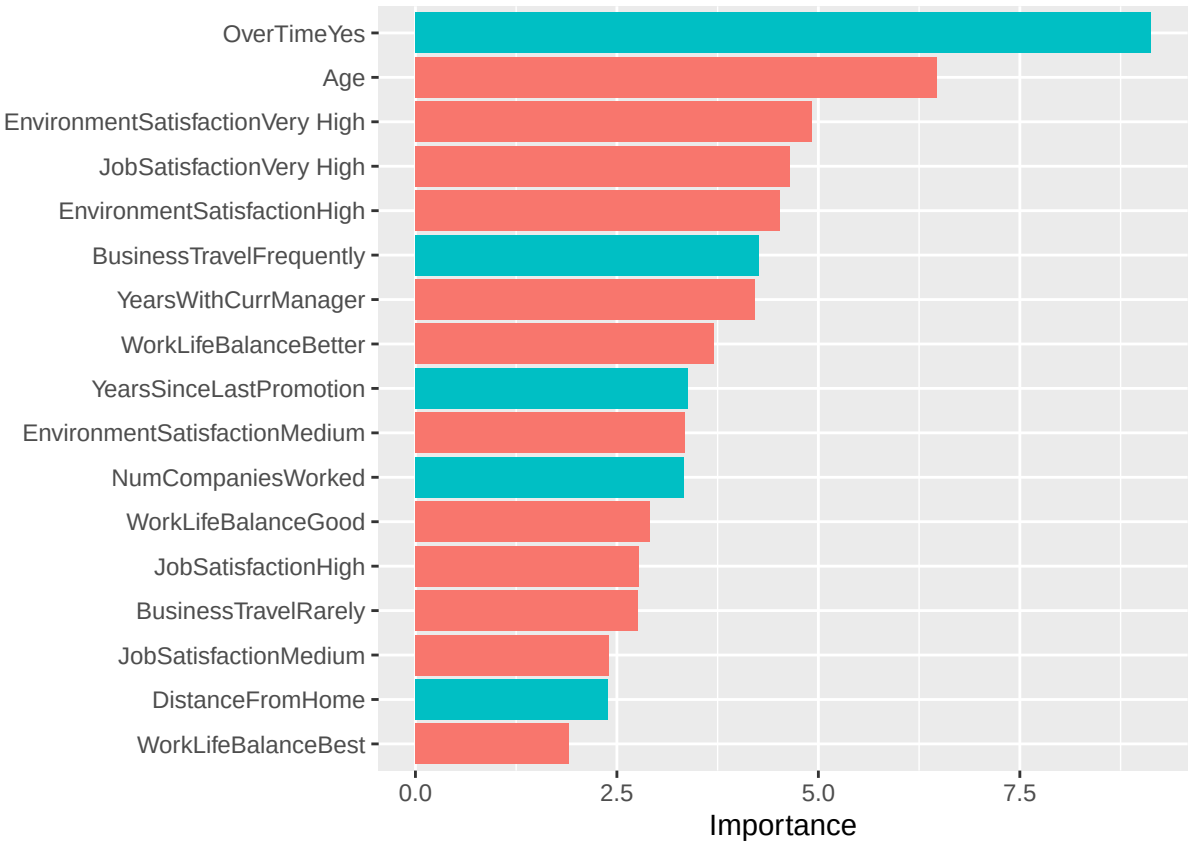


In the above plots we can observe the most important variables (Variable Importance) to predict employment attrition according to the absolute value of the z-statistic for each coefficient in the dataset. Moreover the importance of independent variables are colored to indicate increasing (blue) or decreasing (red) risk of employee attrition. We again observe that OverTime seems to be highly correlated with employee attrition in this data set. Moreover, EnviromentSatisfaction and JobSatisfaction seem to be also be critical, which would make sense since we are talking about employment attrition.

Based on the exploratory data analysis, the previous section as well as the above variable importance scores we are trying to fit a better multiple logistic regression model.

```
glm.model.2 <- glm(Attrition ~ OverTime + EnvironmentSatisfaction + NumCompaniesWorked +
                     JobSatisfaction + BusinessTravel + DistanceFromHome +
                     WorkLifeBalance + Age + YearsWithCurrManager + YearsSinceLastPromotion,
                   family = "binomial", data = emp_attrition_train)
```

The new model seems to have quite a good fit with all independent variables having significant p-values.



The new variable importance plot seems to have ranked Age a lot higher than before while overtime still seems to remain a large main effect. Thus, there might be interactions in these variable and we should start developing the model.

7.4 Model Development

The code for the model development is hidden since it is to large for this paper.

By using the drop1 function we have added and remove significant interactions. Finally we end up with significant interaction inclusion of JobSatifaction with Age, Number of Companies Worked for, Work Life Balance and Business Travel Frequency.

```
final.glm <- glm(Attrition ~ EnvironmentSatisfaction + NumCompaniesWorked + JobSatisfaction +
    BusinessTravel + DistanceFromHome + OverTime + YearsAtCompany +
    PercentSalaryHike + WorkLifeBalance + Age + YearsWithCurrManager +
    YearsSinceLastPromotion +
    JobSatisfaction:Age + JobSatisfaction:NumCompaniesWorked +
    JobSatisfaction:WorkLifeBalance + JobSatisfaction:BusinessTravel, family = "binomial", data = emp_a
# exp(coef(final.glm))
```

Interpretation of the Logistic Regression

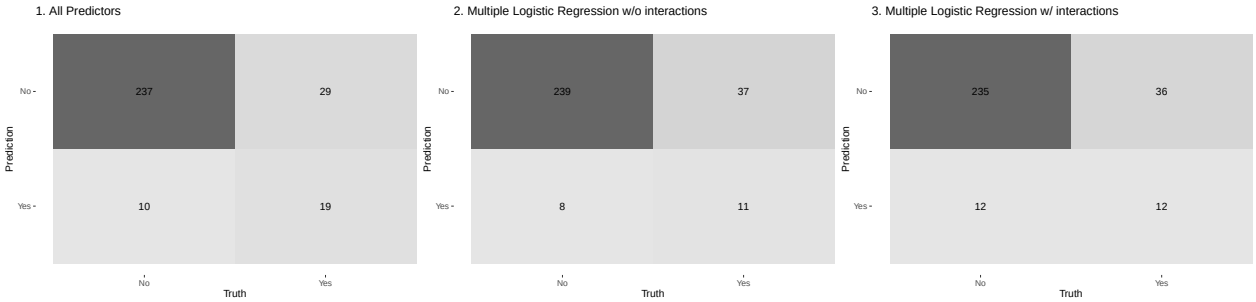
The odds of someone leaving the company

- with overtime are about ~5.6 times higher than the odds for no overtime
- with a 'very high' environment satisfaction are lower than the ones of low environment satisfaction by ~0.31 times.
- with a 'high' environment satisfaction are lower than the ones of low environment satisfaction by ~0.37 times.
- with a 'medium' environment satisfaction are lower than the ones of low environment satisfaction by ~0.38 times.
- are multiplied by 1.19 i.e increasing for each additional company the employee has worked for
- with a 'very high' job satisfaction are lower than the ones of low job satisfaction by ~0.009 times.
- with a 'high' job satisfaction are lower than the ones of low job satisfaction by ~0.12 times.
- with a 'medium' job satisfaction are lower than the ones of low job satisfaction by ~0.04 times.
- having to travel for work 'rarely' increases the risk by ~1.38 times
- having to travel for work 'frequently' increases the risk by ~2.04 times
- are multiplied by 1.03 i.e increasing for each unit of distance between work and home
- are multiplied by 0.97 i.e slightly decreasing for each additional year the employee has worked at the company
- are multiplied by 0.99 i.e slightly decreasing for each percentage in salary hike an employee has received
- with a 'good' work life balance score are lower than the ones of 'bad' work life balance score by ~0.11 times.
- with a 'better' work life balance score are lower than the ones of 'bad' work life balance score by ~0.11 times.
- with a 'best' work life balance score are lower than the ones of 'bad' work life balance score by ~0.12 times.
- are multiplied by 0.9 i.e decreasing for each additional year of age an employee has
- are multiplied by 0.86 i.e decreasing for each additional year an employee has worked for the same manager
- are multiplied by 1.17 i.e increasing for each year an employee has not received a promotion

Summary

Overtime seems to increase the risk of attrition by almost six times as much and is by far the most highly critical attribute on whether an employee continues to stay at the company or not. It does not seem to matter to much how well the *job environment* score is as long as it is above 'low' as they almost equally decrease the risk of someone leaving the company. The amount of *business traveling* an employee has to do seems to play an important role as well as the risk of someone quitting the company increases by two times if said person has frequently travel for work. *Age* seems to play a role as well as young employee seem to leave the company more often than old employees. Employee seem to leave the company less often if they are not bound to re-organisations i.e. have the same *manager* for an extended period of time. For each year an employee has not received a *promotion* the risk of them leaving increases.

7.5 Confusion Matrix



1. When looking at the first plot using cross validation & including all predictors we get the following scores.

   Sensitivity = True Negatives / (True Negatives + False Positives)

   - i.e. 237 / (10 + 237) = 0.95
   - 95% of the people not leaving the company were correctly identified by the Logistic Regression model.

   Specificity = True Positives / (True Positives + False Negatives)

   - i.e. 19 / (29 + 19) = 0.4
   - 40% of the people leaving the company were correctly identified by the Logistic Regression model.

2. Without interactions using a much simpler model we actually get a better specificity, but can only predict 23% of the people who are actually leaving the company. In other words the True Positive Rate which we are looking for is significantly worse.

   Sensitivity = 239 / (8 + 239) = 0.97

   - 97% of the people not leaving the company were correctly identified by the Logistic Regression model.

   Specificity = 11 / (37 + 11) = 0.23

   - 23% of the people leaving the company were correctly identified by the Logistic Regression model.

3. When including the interactions we are able to predict the True Positive Rate a bit better however we sacrifice some of the specificity

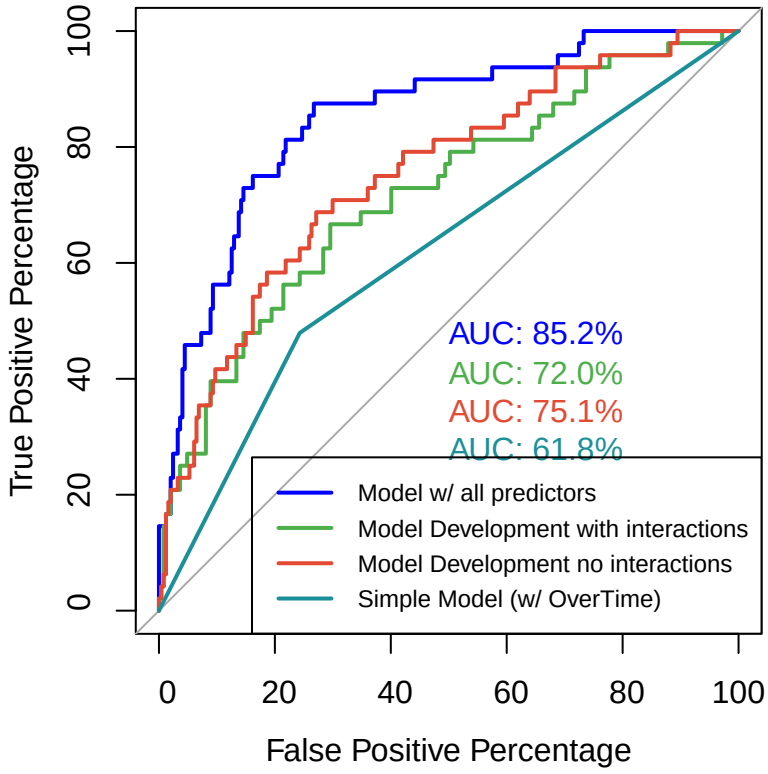
   Sensitivity = 235 / (12 + 235) = 0.95

   - 95% of the people not leaving the company were correctly identified by the Logistic Regression model.

   Specificity = 12 / (36 + 12) = 0.25

   - 25% of the people leaving the company were correctly identified by the Logistic Regression model.

7.6 Receiver Operating Characteristics



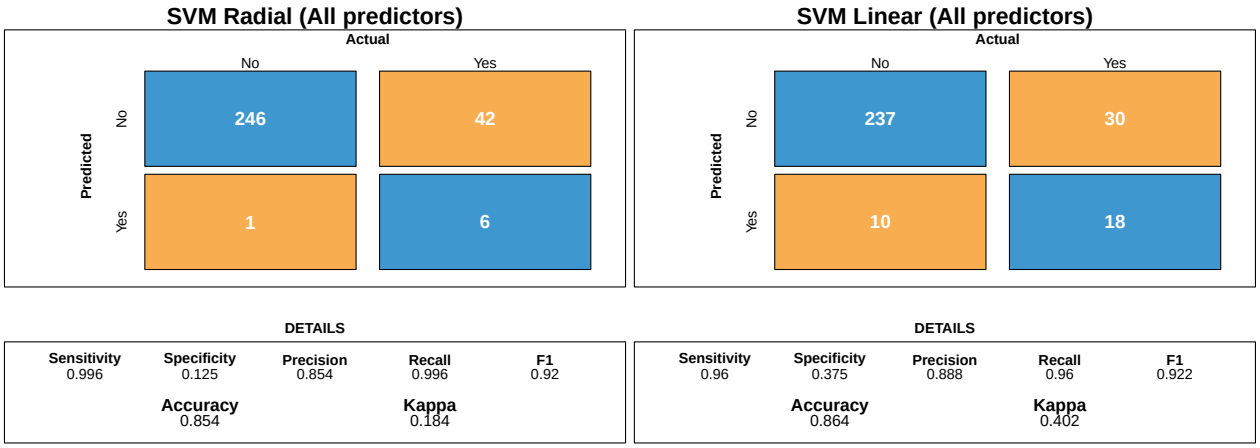
The above plot illustrates how the True Positive Rate (Sensitivity) behaves in relation with the False Positive Rate (1-Specificity). In this paper we want to maximize the amount of correct classifications of people leaving the company i.e. True Positive Rate.

Thus we can take away from the above plot that including the interactions does not make to much sense since with a very similar False Positive Rate (around 25-30%) with a much simpler model (12 independent variables & 4 interactions vs 31 total independent variables).

8 Support Vector Machine

8.1 Model Training

With same training testing data set from above models we will train two different type of SVM model Linear and Radia and observe its performance.



Here we can observe that both model has relatively poor specificity compare to its accuracy. This means that model is trained biased to predict that given employee does not have Attrition. This is due to unbalanced label of data set where most of our data set were employees without Attrition.

Depending on the objective of the business goal this model maybe good enough, but our goal is to minimize the cost happening due to Attrition so we may need a model that has reasonable specificity.

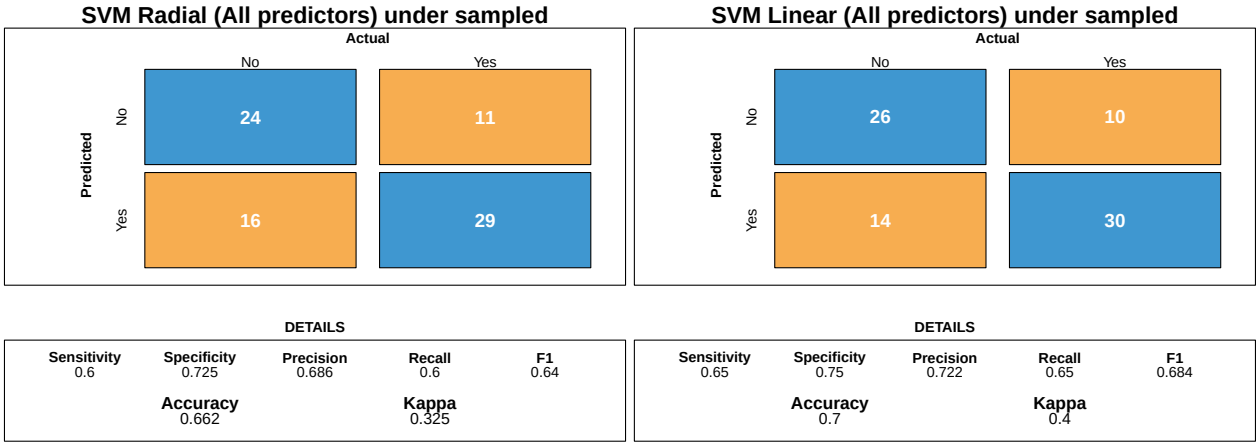
8.2 Under sampling

In order to avoid biased training we will under sample the data set to have equal amount of employee with Attrition and without.

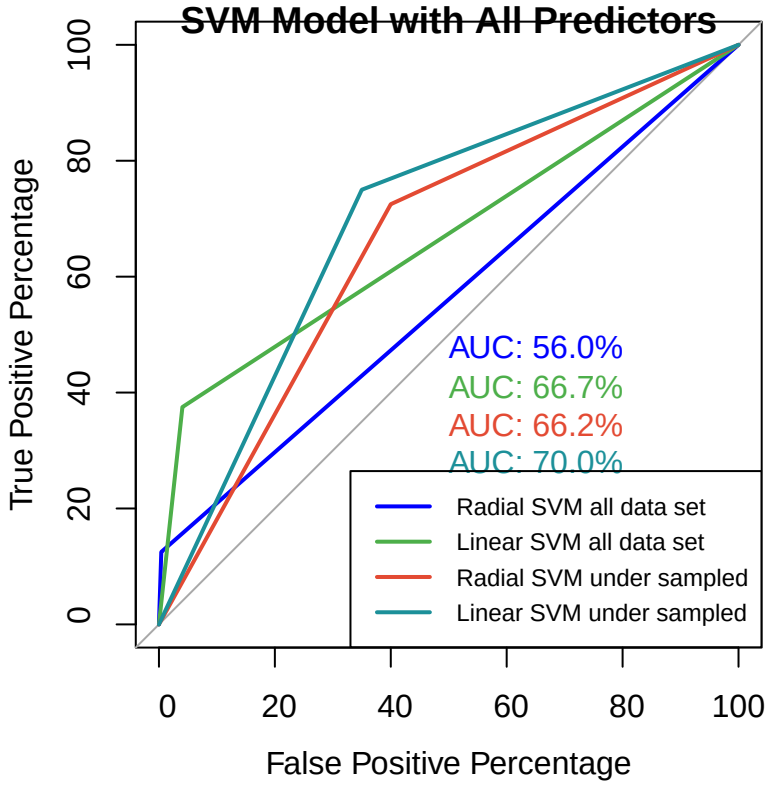
```
set.seed(111)
emp_undersample<- emp_attrition %>% group_by(Attrition) %>% slice_sample(n=200)

emp_attrition_split.under <- initial_split(emp_undersample, prop = 0.80, strata = "Attrition")
emp_attrition_train.under <- training(emp_attrition_split.under)
emp_attrition_test.under  <- testing(emp_attrition_split.under)
emp_undersample %>% count(Attrition)
```

```
## # A tibble: 2 x 2
## # Groups:   Attrition [2]
##   Attrition     n
##   <fct>     <int>
## 1 No          200
## 2 Yes         200
```

Here we can observe that under sampled dataset has equal amount of 'Yes' and 'No'. With this new data set we will train SVM and observe its performance.



After under sampling the data to have balanced label the perforamnce of both model changed. The accuracy of both model decreased but specificity increased. This indicates that this model is more capable of predicting employee who may have Attrition.



By observing the ROC curve and AUC value of each models we can see Linear SVM under sampled model shows the highest AUC value.

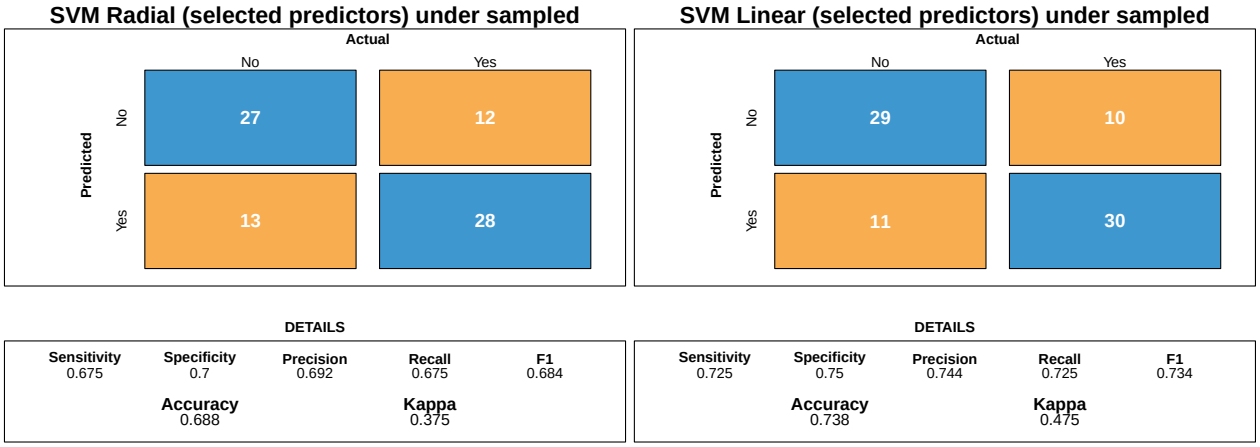
8.3 Variable Selection

In order to compare with other models with how selection of predictor/variable impacts the model performance we trained SVM with selected variables.

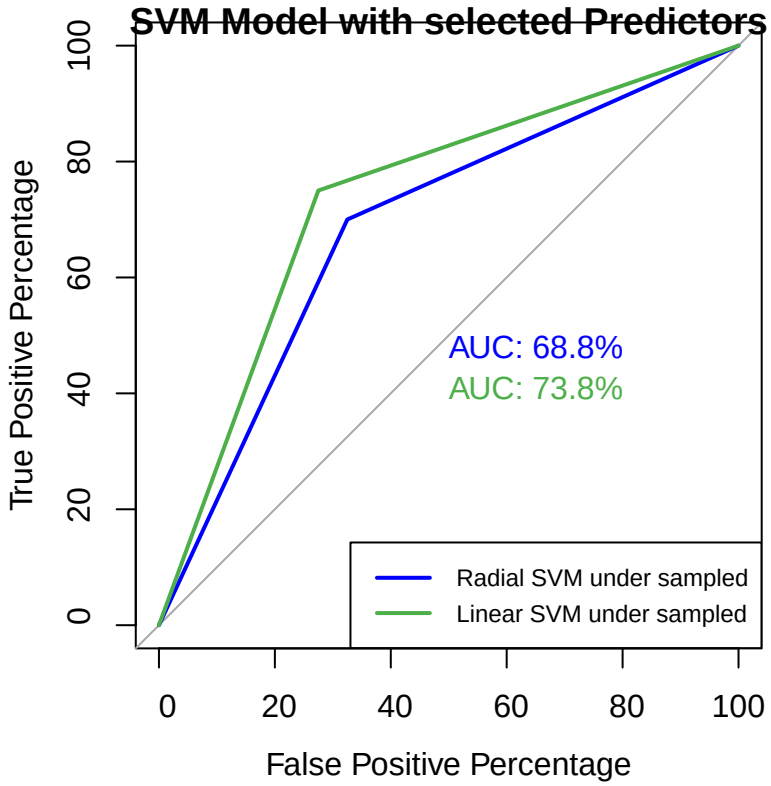
Chosen variables are shown below.

(EnvironmentSatisfaction, NumCompaniesWorked, JobSatisfaction, BusinessTravel, DistanceFromHome, OverTime, YearsAtCompany, PercentSalaryHike, WorkLifeBalance, Age, YearsWithCurrManager, YearsS-

inceLastPromotion)



There doesn't seem to be a significant difference between these two model to the models with all predictors. They both have similar accuracy and specificity, so we can assume that SVM models does not get highly impacted by the variable selection as GLM models did.

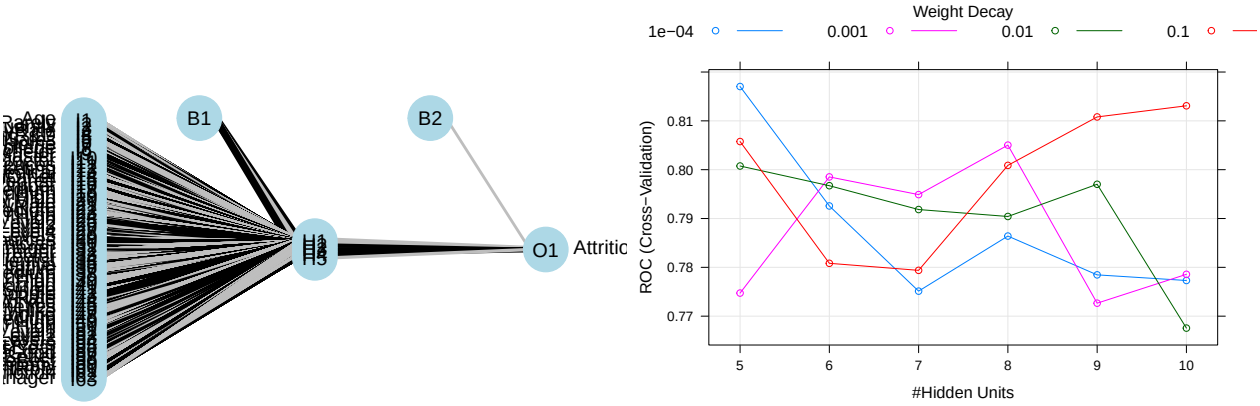


By observing the ROC curve and AUC value of each models with selected variables we observe the AUC value is slightly higher but this may not be significant due to small size of data set.

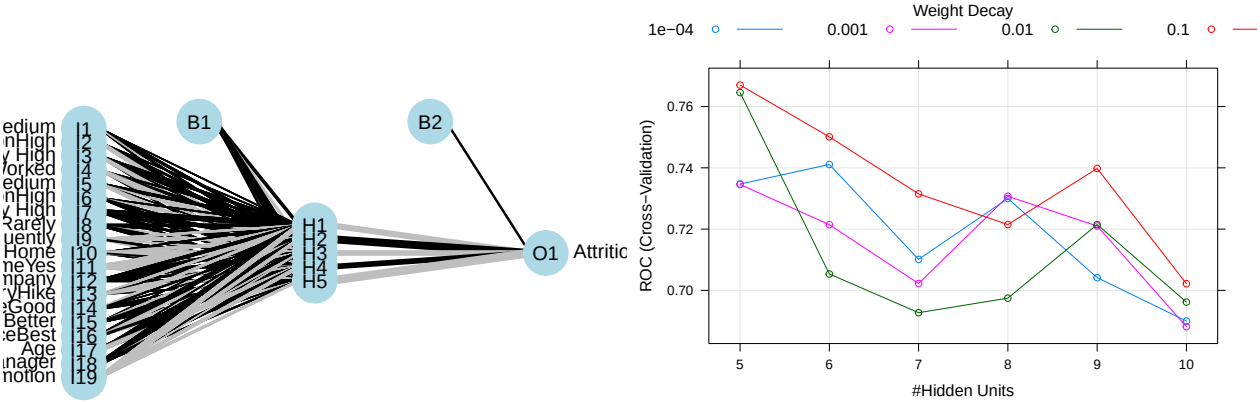
9 Neural Network

9.1 Model Training

We trained a simple neural network by tuning the hidden layer size and decay.

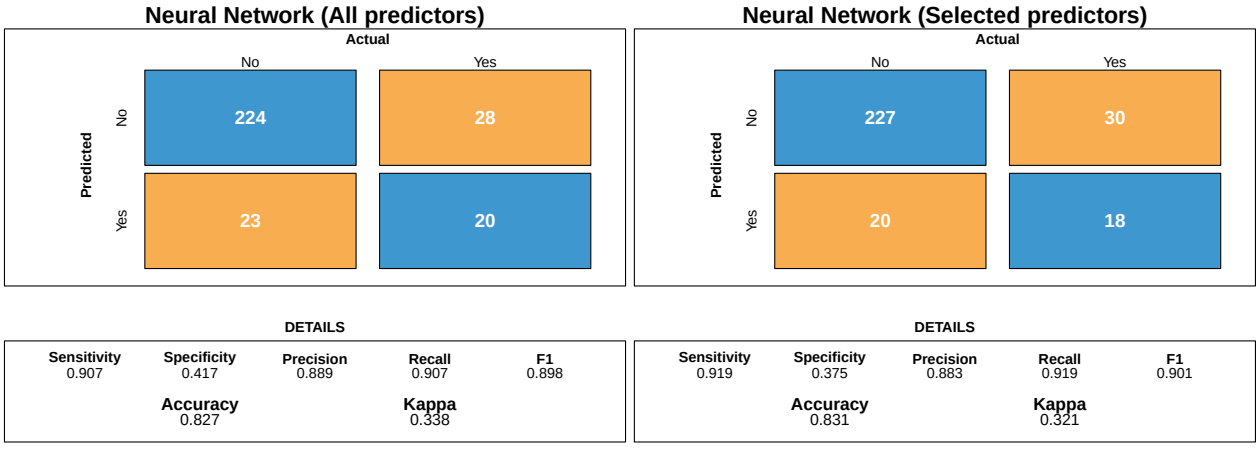


Above figure shows the neural network model with input layer, hidden layer and the output layer.

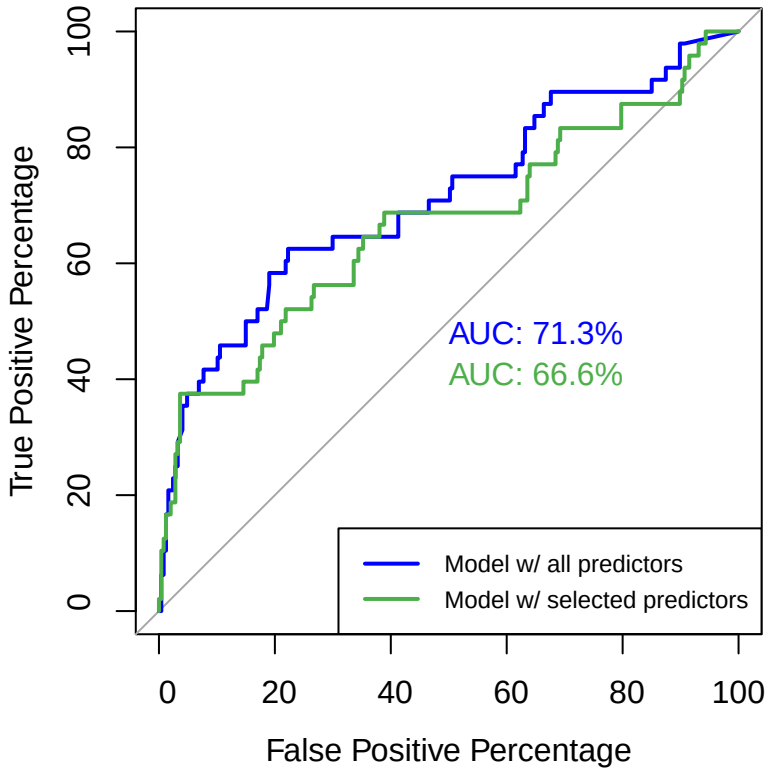


We train another model with only selected predictors. Above figure shows the neural network model with input layer, hidden layer and the output layer.

Same as previously parameter tuning shows the optimized value.



Above is the confusion matrix of the neural network model. As we can see still the specificity of the model is quite low. Since Neural Networks final output is a probability, we can adjust our cutoff value to have higher specificity.



By observing the ROC curve we can estimate the performance of the model better than just observing the accuracy of the model as our business goal is not just to have good accuracy but also predict employees with Attrition. Comparing AUC value among model will tell us which model performs better.

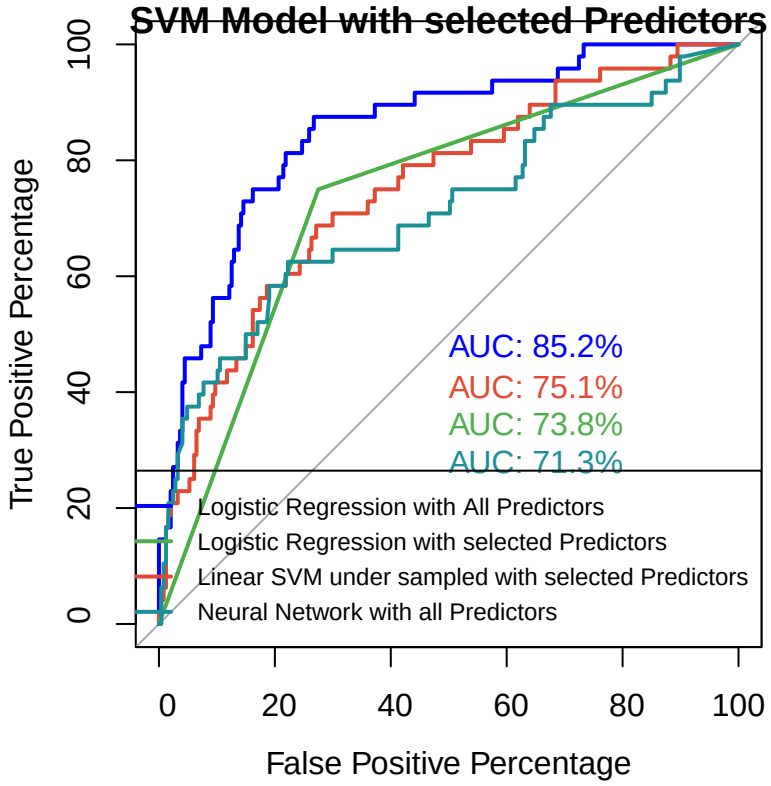
10 Conclusion

In conclusion of the performed models, starting with the Linear Model and the General Additive Models, the latter perfomed slightly better due to included smooth terms, if measured by the Adjusted R-squared values. Similar effects were displayed for the fading of influence of the categorical variables Attrition, EnvironmentSatisfaction and WorkLifeBalance on the dependent variable YearsAtCompany for both approaches, when fitted to the smaller test dataset. Taking into account that the Linear Model allows to simply interpret the influence of the predictor variables, the Final Linear Model should be preffered to gain insights on the Attrition dataset compared to the more complex GAM Model.

Of the observed significant influences in Final Linear Model the following seem crucial in our perspective:

- YearsWithCurrManager: The dependent variable, YearsAtCompany, increases by 0.75% if the independent variable increases by 1%. It is not surprising that people seem to prefer consictency in their direct management but this has to be kept in mind in regard of promotions but also possible restructuring.
- YearsSinceLastPromotion: The dependent variable, YearsAtCompany, increases by 5.50% if the independent variable increases by 1 year. People who might be waiting/expecting a promotion are likely to work longer for a company. Therefore, a promotion system should be setup on trustworthy policy with transparent rules and derived individual goals, to allow people to work towards a promotion.
- WorkLifeBalance: The dependent variable, YearsAtCompany, increases by 6.11% respectively by 8.98% if the balance is "good" or even "best," compared to the reference level "better." The better the WorkLifeBalance of employees, the longer they seem to keep working for the company. This has to be kept in mind when for example ordering people back to the headquarters in post-pandemic times.

Going forward, we compared three different category of models (Logistic Regression, SVM, and Neural Network). As our business goal was to have a robust model that can predict both employees with Attrition and without Attrition well, we used AUC value to compare the performance of the model. The result shows that Logistic Regression model with All Predictors had the highest AUC values.



Session Information

```
sessionInfo()
```

```
## R version 4.1.2 (2021-11-01)
## Platform: aarch64-apple-darwin20 (64-bit)
## Running under: macOS Monterey 12.3.1
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.1-arm64/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.1-arm64/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] magicfor_0.1.0       multcomp_1.4-19      TH.data_1.1-1
##  [4] survival_3.2-13      mvtnorm_1.1-3        NeuralNetTools_1.5.3
##  [7] nnet_7.3-16          pROC_1.18.0          arm_1.12-2
## [10] lme4_1.1-27.1        Matrix_1.3-4         MASS_7.3-54
## [13] yardstick_0.0.9      forcats_0.5.1        stringr_1.4.0
## [16] purrr_0.3.4          readr_2.1.2          tidyr_1.2.0
## [19] tibble_3.1.6         tidyverse_1.3.1      rsample_0.1.1
## [22] vip_0.3.2            caret_6.0-92         lattice_0.20-45
## [25] mgcv_1.8-38          nlme_3.1-155         kableExtra_1.3.4
## [28] papeR_1.0-5          xtable_1.8-4         car_3.0-12
## [31] carData_3.0-5        ggplot2_3.3.6        dplyr_1.0.9
## [34] bookdown_0.24
##
## loaded via a namespace (and not attached):
##  [1] minqa_1.2.4          colorspace_2.0-2     ellipsis_0.3.2
##  [4] class_7.3-19         fs_1.5.2             proxy_0.4-26
##  [7] rstudioapi_0.13      farver_2.1.0         listenv_0.8.0
## [10] furrr_0.3.0          prodlim_2019.11.13   fansi_1.0.2
## [13] lubridate_1.8.0      xml2_1.3.3           codetools_0.2-18
## [16] splines_4.1.2        knitr_1.37           jsonlite_1.7.3
## [19] nloptr_2.0.0         broom_0.7.12         kernlab_0.9-30
## [22] dbplyr_2.1.1         compiler_4.1.2       httr_1.4.2
## [25] backports_1.4.1      assertthat_0.2.1     fastmap_1.1.0
## [28] cli_3.2.0            htmltools_0.5.2      tools_4.1.2
## [31] coda_0.19-4          gtable_0.3.0         glue_1.6.1
## [34] reshape2_1.4.4       gmodels_2.18.1       Rcpp_1.0.8
## [37] cellranger_1.1.0     vctrs_0.4.1          gdata_2.18.0
## [40] svglite_2.1.0        iterators_1.0.13     timeDate_3043.102
## [43] gower_0.2.2          xfun_0.29            globals_0.14.0
## [46] rvest_1.0.2          lifecycle_1.0.1      gtools_3.9.2
## [49] future_1.25.0        zoo_1.8-9            scales_1.1.1
## [52] ipred_0.9-12         hms_1.1.1            sandwich_3.0-1
## [55] parallel_4.1.2       yaml_2.3.4           gridExtra_2.3
## [58] rpart_4.1-15         stringi_1.7.6        highr_0.9
## [61] foreach_1.5.1        e1071_1.7-9          boot_1.3-28
```

```
## [64] hardhat_0.2.0        lava_1.6.10          rlang_1.0.2
## [67] pkgconfig_2.0.3      systemfonts_1.0.3    evaluate_0.14
## [70] labeling_0.4.2       recipes_0.2.0        tidyselect_1.1.1
## [73] parallelly_1.30.0    plyr_1.8.6           magrittr_2.0.2
## [76] R6_2.5.1             generics_0.1.2       DBI_1.1.2
## [79] pillar_1.7.0         haven_2.4.3          withr_2.4.3
## [82] abind_1.4-5          future.apply_1.8.1   modelr_0.1.8
## [85] crayon_1.5.0         utf8_1.2.2           tzdb_0.2.0
## [88] rmarkdown_2.11       grid_4.1.2           readxl_1.3.1
## [91] data.table_1.14.2    ModelMetrics_1.2.2.2 reprex_2.0.1
## [94] digest_0.6.29        webshot_0.5.2        stats4_4.1.2
## [97] munsell_0.5.0        viridisLite_0.4.0
```

References

IBM. (2019). *IBM HR analytics employee attrition & performance*. https://github.com/IBM/employee-attrition-aif360/blob/master/data/emp_attrition.csv

Swaminathan, S., & Hagarty, R. (2020). *IBM HR analytics employee attrition & performance* (2nd ed.). IBM. https://developer.ibm.com/patterns/data-science-life-cycle-in-action-to-solve-employee-attrition-problem/