Employee attrition

Petra Kaferle Devisschere,Alix Sarrazin and Yu-Hsuan Ting

29/11/2019

1. Introduction

Employee turnover is an important issue for each company as filling the vacancies is costly procedure, financially and timewise, especially for critical positions. On the other hand, planned departures gives employers the opportunities to revise their organisational structure and potentially close down the positions that bring less benefits to the organisation. The same approach has been used to look for biases in employee retention in specific working environments (e.g. Survival Analysis of Faculty Retention and Promotion in the Social Sciences by Gender, Box-Steffensmeier et al. 2015). In this project we explored synthetic dataset, simulating employee data from HR database. First, we formatted it and constructed additional variables. Then, we applied different methodologies to find the model that would best describe the relationship between available attributes and employee withdrawal.

1.1 Dataset description

We downloaded data in October 2019 from: http://rpubs.com/rhuebner/HRCodebook-13. In total, there are 310 observations with 35 variables. Below is the description of each variable.

Data Dictionary

Feature	**Description**	**DataType**
Employee Name	Employee's full name	Text
EmpID	Employee ID is unique to each employee	Text
MarriedID	Is the person married (1 or 0 for yes or no)	Binary
MaritalStatusID	Marital status code that matches the text field MaritalDesc	Integer
EmpStatusID	Employment status code that matches text field EmploymentStatus	Integer
DeptID	Department ID code that matches the department the employee works in	Integer
PerfScoreID	Performance Score code that matches the employee's most recent performance score	Integer
FromDiversityJobFairID	Was the employee sourced from the Diversity job fair? 1 or 0 for yes or no	Binary
PayRate	The person's hourly pay rate. All salaries are converted to hourly pay rate	Float
Termd	Has this employee been terminated - 1 or 0	Binary
PositionID	An integer indicating the person's position	Integer
Position	The text name/title of the position the person has	Text
State	The state that the person lives in	Text
Zip	The zip code for the employee	Text
DOB	Date of Birth for the employee	Date
Sex	Sex - M or F	Text
MaritalDesc	The marital status of the person (divorced, single, widowed, separated, etc)	Text

Data Dictionary		
CitizenDesc	Label for whether the person is a Citizen or Eligible NonCitizen	Text
HispanicLatino	Yes or No field for whether the employee is Hispanic/Latino	Text
RaceDesc	Description/text of the race the person identifies with	Text
DateofHire	Date the person was hired	Date
DateofTermination	Date the person was terminated, only populated if, in fact, Termd = 1	Date
TermReason	A text reason / description for why the person was terminated	Text
EmploymentStatus	A description/category of the person's employment status. Anyone currently working full time = Active	Text
Department	Name of the department that the person works in	Text
ManagerName	The name of the person's immediate manager	Text
ManagerID	A unique identifier for each manager.	Integer
RecruitmentSource	The name of the recruitment source where the employee was recruited from	Text
PerformanceScore	Performance Score text/category (Fully Meets, Partially Meets, PIP, Exceeds)	Text
EngagementSurvey	Results from the last engagement survey, managed by our external partner	Float
EmpSatisfaction	A basic satisfaction score between 1 and 5, as reported on a recent employee satisfaction survey	Integer
SpecialProjectsCount	The number of special projects that the employee worked on during the last 6 months	Integer
LastPerformanceReviewDate	The most recent date of the person's last performance review.	Date
DaysLateLast30	The number of times that the employee was late to work during the last 30 days	Integer

1.2 Data preparation

The data pre-processing steps are the following:

- Remove following variables: Employee_Name, EmployeeID, MarriedID, GenderID, epStatusID, DeptID, PerfScoreID, FromDiversityJobFairID, PositionID, State, Zip, HispanicLatino, TermReason, EmploymentStatus, RecruitmentSource, LatestPerformanceReview_Date, DaysLateLast30
- Transform all the data columns from factor to date format
- Determine the last date of the dataset and add 1 to the last date to simulate the probable date of data export ("2017-04-21")
- Change all string-based factors to lower-case to avoid mixed lower/upper case typing
- Create additional groups for the attributes in Table 1. Two different groups were made based on recruitment source, found in the column ProximityRec (proximity recruitment). Value 1 indicates all the categories where recruiters and candidates have some level of personal contact (in contrast to applying online - value 0)

New Group	From column	New Group	From column
AgeGrp	Age	**EmpSatGrp**	EmpSatisfaction
PayGrp	PayRate	**PerfScrGrp**	PerfScoreID
ProximityRec	RecruitmentSource	**IsManager**	position

```
fmt <- "%m/%d/%Y"
data$DateofHire <- as.Date(as.character(data$DateofHire),format=fmt)
data$DateofTermination <- as.Date(as.character(data$DateofTermination),format=fmt)
data$DOB <- as.Date(as.character(data$DOB),format = fmt)
endtime <- max(data$DateofHire,data$DateofTermination,na.rm = TRUE) + 1

is_manager<-(str_detect(data$Position,"Manager")|str_detect(data$Position,"Director")|
               str_detect(data$Position,"CIO")|str_detect(data$Position,"CTO"))

data<-mutate(data,Duration=ifelse(is.na(DateofTermination),(endtime - DateofHire)/365.25,
                     (DateofTermination - DateofHire)/365.25),
             Age=as.numeric(floor((endtime-DOB)/365.25)),
             IsManager=ifelse(is_manager,"is manager","not manager"))

data$AgeGrp<-cut(data$Age,breaks = c(20,30,40,50,60,70),right = FALSE,
                 labels = c("20-29","30-39","40-49","50-59","60-69"))
data$PayGrp<-cut(data$PayRate,breaks = c(10,20,30,40,50,60,90),right = FALSE,
                 labels = c("10-19","20-29","30-39","40-49","50-59",">60"))
data$PerfScrGrp<-cut(data$PerfScoreID,breaks = c(1,2.99,3.01,5),right = FALSE,
                   labels = c("low","medium","high"))
data$EmpSatGrp<-cut(as.numeric(data$EmpSatisfaction),breaks = c(1,2.99,3.01,6),
                  right = FALSE,labels = c("low","medium","high"))

data$Sex  <-  factor(tolower(data$Sex))
data$PositionID <- factor(tolower(data$PositionID))
data$ManagerID <- factor(tolower(data$ManagerID))
data$MaritalDesc <- factor(tolower(data$MaritalDesc))
data$Department <- factor(tolower(data$Department))
data$ManagerName <- factor(tolower(data$ManagerName))
data$RecruitmentSource <- factor(tolower(data$RecruitmentSource))
data$PerformanceScore <- factor(data$PerformanceScore)
data$EmpSatisfaction <- factor(data$EmpSatisfaction)
data$IsManager<-factor(data$IsManager)

prox <- c('company intranet - partner','diversity job fair','employee referral',
          'information session','on-campus recruiting','professional society',
          'social networks - facebook twitter etc','vendor referral','word of mouth' )
data <- mutate(data,ProximityRec = ifelse(tolower(RecruitmentSource) %in% prox,1,0))
data$ProximityRec=factor(data$ProximityRec)
```

To reduce the size of the dataset we will be manipulating, we selected only specific columns.

```
df <-subset(data,select=c(Sex,Age,AgeGrp,MaritalDesc,Position,IsManager,Department,
            ManagerName,PayRate,PayGrp,SpecialProjectsCount,PerfScrGrp,EngagementSurvey,
            EmpSatisfaction,EmpSatGrp,RecruitmentSource,ProximityRec,Duration,Termd))
```

The summary of working dataset shows that we have no missing values, nor extreme outliers for any variable.

```
##  Sex          Age          AgeGrp       MaritalDesc                      Position
##  f :177   Min.   :24.00   20-29: 55   divorced : 30   Production Technician I :136
##  m :133   1st Qu.:30.00   30-39:143   married  :123   Production Technician II: 57
##           Median :36.00   40-49: 79   separated: 12   Area Sales Manager      : 27
##           Mean   :37.71   50-59: 23   single   :137   Production Manager      : 14
##           3rd Qu.:43.00   60-69: 10   widowed  :  8   Software Engineer       :  9
```

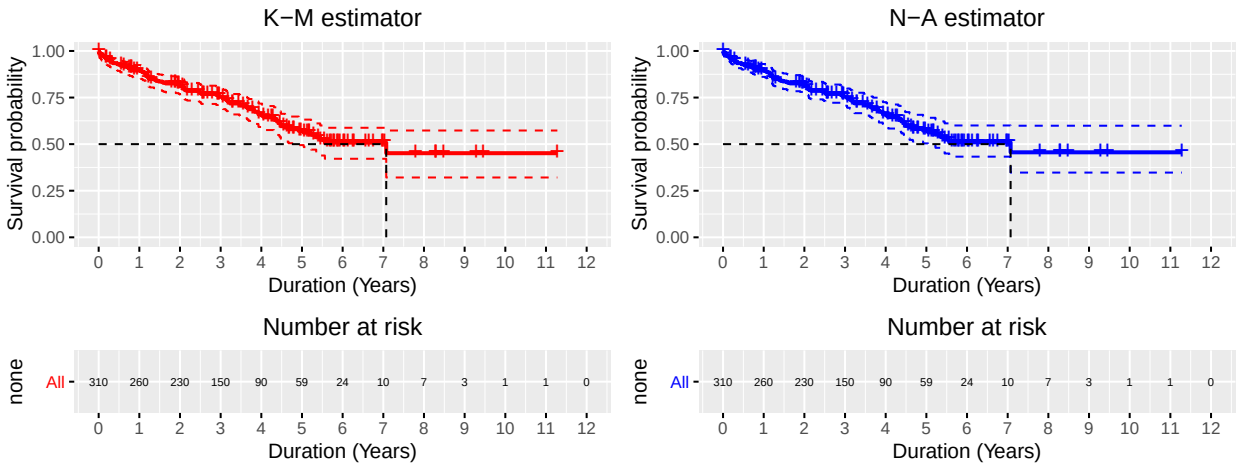
```
##           Max.   :66.00                                IT Support              :  8
##                                                        (Other)                 : 59
##       IsManager                  Department          ManagerName     PayRate        PayGrp
##  is manager : 56   admin offices       : 10   elijiah gray  : 22   Min.   :14.00   10-19: 72
##  not manager:254   executive office    :  1   kelley spirea : 22   1st Qu.:20.00   20-29:134
##                    it/is               : 50   kissy sullivan: 22   Median :24.00   30-39: 12
##                    production          :208   michael albert: 22   Mean   :31.28   40-49: 24
##                    sales               : 31   amy dunn      : 21   3rd Qu.:45.31   50-59: 56
##                    software engineering: 10   brannon miller: 21   Max.   :80.00   >60  : 12
##                                               (Other)       :180
##  SpecialProjectsCount  PerfScrGrp  EngagementSurvey EmpSatisfaction  EmpSatGrp
##  Min.   :0.00         low   : 30   Min.   :1.030    1:  2           low   : 11
##  1st Qu.:0.00         medium:243   1st Qu.:2.083    2:  9           medium:108
##  Median :0.00         high  : 37   Median :3.470    3:108           high  :191
##  Mean   :1.21                      Mean   :3.332    4: 93
##  3rd Qu.:0.00                      3rd Qu.:4.520    5: 98
##  Max.   :8.00                      Max.   :5.000
##
##                          RecruitmentSource ProximityRec    Duration            Termd
##  employee referral                 : 31     0:174        Min.   : 0.002738   Min.   :0.0000
##  diversity job fair                : 29     1:136        1st Qu.: 1.967830   1st Qu.:0.0000
##  search engine - google bing yahoo: 25                   Median : 2.943190   Median :0.0000
##  monster.com                       : 24                  Mean   : 3.148104   Mean   :0.3323
##  pay per click - google            : 21                  3rd Qu.: 4.453799   3rd Qu.:1.0000
##  professional society              : 20                  Max.   :11.279945   Max.   :1.0000
##  (Other)                           :160
```

2. Data and methods

2.1 Descriptive analysis with Kaplan-Meier estimator

First, we looked at the global survival curve, without taking censoring into account. We estimated the expected stay in the company and compared Kaplan-Meier and Nelson-AAlen estimators. Further on, we explored the dataset by comparing differences between subgroups of the following variables: Sex, PayGrp, Manager and Department. Is any of them significantly contributing to employee turnover?

```
y<-with(df,Surv(Duration ,Termd))
km<-survfit(y ~ 1,data = df,conf.type = "log-log")
na<-survfit(y ~ 1,data=df,type="fh")

splots <- list()
splots[[1]]<- ggsurvplot(km,data=df,risk.table =T,palette= 'red',censor=T,
                         conf.int.style="step",xlab='Duration (Years)',break.time.by = 1,
                         surv.median.line='hv',
                         risk.table.fontsize = 2,fontsize = 2,risk.table.height = 0.35,
                         legend='none',legend.title="none",title='K-M estimator',
                         ggtheme = theme(plot.title = element_text(hjust = 0.5)))
splots[[2]]<- ggsurvplot(na,data=df,risk.table =T,palette ='blue',censor=T,
                          conf.int.style="step",xlab='Duration (Years)',break.time.by = 1,
                          surv.median.line='hv',
                          risk.table.fontsize = 2,fontsize = 2,risk.table.height = 0.35,
                          legend='none',legend.title="none", title='N-A estimator',
                          ggtheme = theme(plot.title = element_text(hjust = 0.5)))
arrange_ggsurvplots(splots,print = TRUE,ncol = 2,nrow = 1)
```



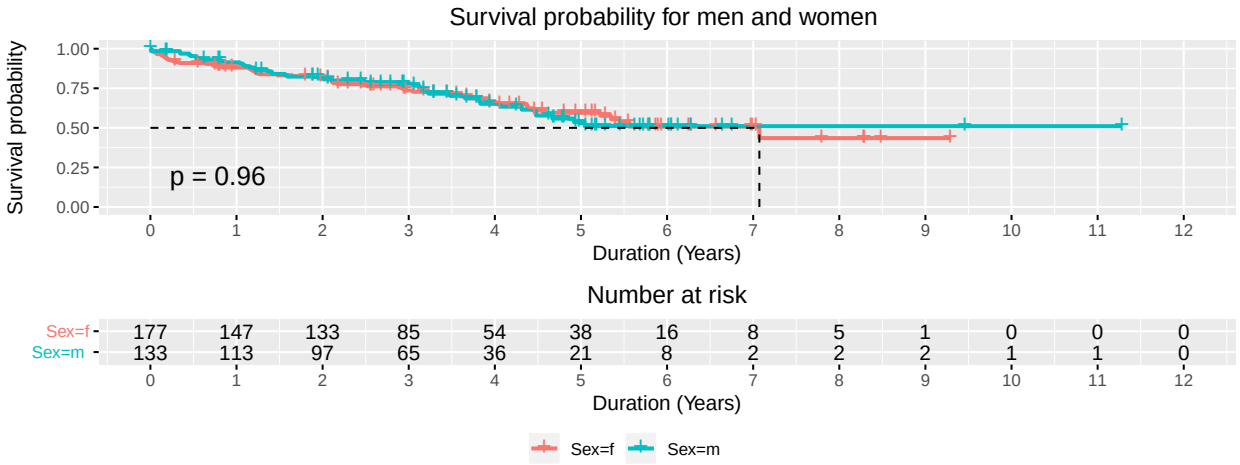
Estimator	N	Events	Median	0.95 LCL	0.95 UCL
K-M	310.00	103.00	7.07	4.92	NA
N-A	310.00	103.00	7.07	5.04	NA

The results show that both metrics give roughly the same results. Expected stay in this company is about 7 years. Lower limits of confidence interval differ, but remain very similar. The upper CI limit cannot be estimated, since it never drops below 50%. It can also be noticed that all departures happen before employees reach 7.07 years of tenure.

2.1.1 Sex

Are either men or women more likely to leave?

```
gen<- survfit(y ~ Sex,data=df,conf.type="log-log")
ggsurvplot(gen,data=df,risk.table =T,censor=T,xlab='Duration (Years)',break.time.by = 1,
           legend="bottom",legend.title="",risk.table.fontsize = 4,pval=T,
           title = "Survival probability for men and women ",
           risk.table.height = 0.35,surv.median.line='hv',
           ggtheme = theme(plot.title = element_text(hjust = 0.5)))
```

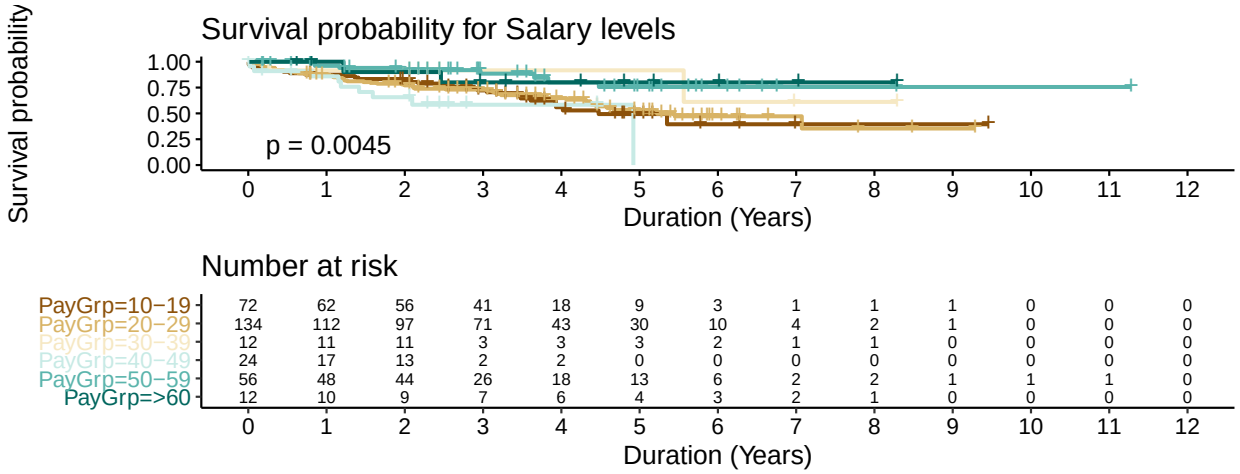


There is no statistical difference between genders, which is evident from the overlap of survival curves and the matching p-value.

2.1.2 Salary level

Are badly paid employees more likely to leave?

```
pal<- brewer.pal(6,"BrBG")
pay<- survfit(y ~ PayGrp,data=df)
ggsurvplot(pay,data=df,risk.table =T,palette =pal,censor=T,xlab='Duration (Years)',
break.time.by = 1,legend='none',pval=T,legend.title="",
title = "Survival probability for Salary levels",
risk.table.fontsize = 3,risk.table.height = 0.50)
```



There is a clear trend between pay group and survivability. The low global p-value tells us that we can reject null hypothesis - that all beta coefficients are equal to zero - but it doesn't tell us which one of them are different. To find it out, we can use pairwise_survdiff to compare all pairs of groups.

```
t1=pairwise_survdiff(Surv(Duration,Termd) ~PayGrp,data=df)
pander(t1, style ='rmarkdown')
```

- **method**: Log-Rank test
- **data.name**: df and PayGrp
- **p.value**:

	10-19	20-29	30-39	40-49	50-59
20-29	0.8735	NA	NA	NA	NA
30-39	0.2793	0.2793	NA	NA	NA
40-49	0.2531	0.2603	0.09906	NA	NA
50-59	0.02009	0.02009	0.8735	0.0064	NA
>60	0.2531	0.2531	0.8735	0.1758	0.9533

- **p.adjust.method**: BH

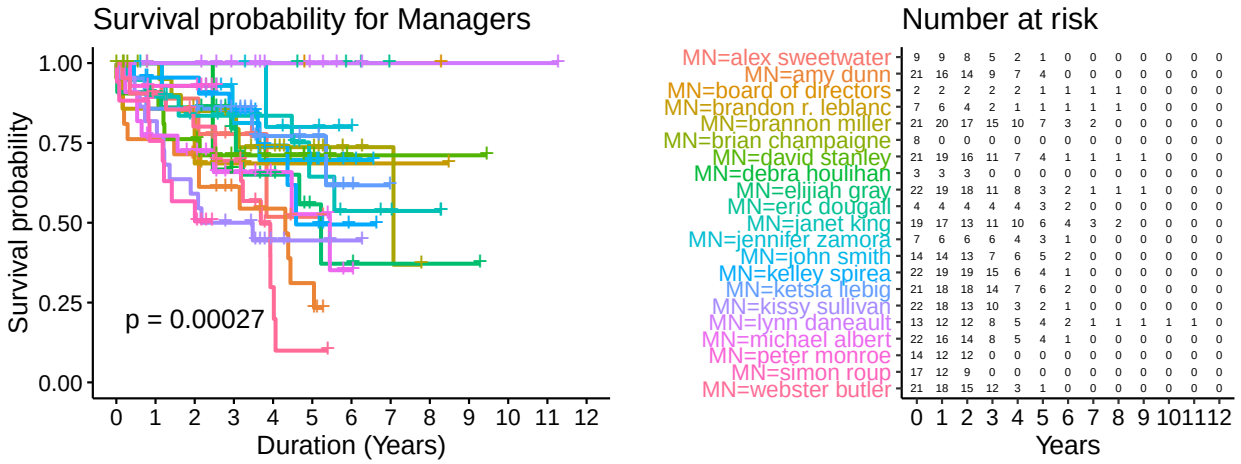
Statistically significant differences can be observed only between individual groups. From the plot above, it was expected to detect some amount of different behaviour between group 10-19 with 30-39, 50-59 and >60. Indeed, the p-values drop significantly after the comparison 10-19 with 30-39, but not under 0.05. Moreover, groups 10-19 and 20-29 display almost the same p-values in the comparisons and have overlaping survival curves. Therefore, we decided to treat them as one group.

The results suggest that salary level influences the duration of stay to certain extend, but the critical condition is if being paid below or above 30.

2.1.3 Manager

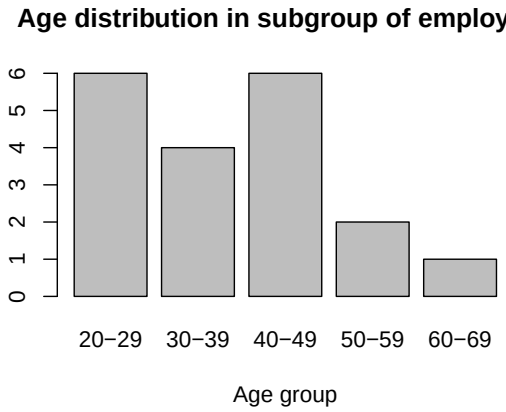
Is there a manager with higher turnover?

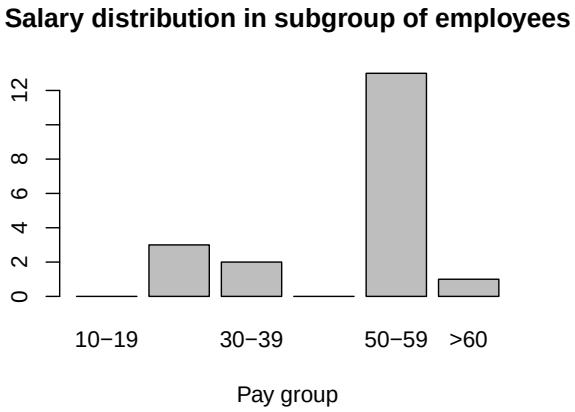
```
colnames(df)[8] <- "MN"
mng<-survfit(y ~ MN,data=df)
splots <- list()
splots[[1]]<-ggsurvplot(mng,data=df,risk.table =F,censor=T,pval=T,
           xlab='Duration (Years)',break.time.by = 1,legend='none',legend.title="",title = "Survival pr
splots[[2]]<-ggsurvplot(mng,data=df,risk.table =T,censor=T,
           xlab='Years',break.time.by = 1,legend='none',
           legend.title="",risk.table.fontsize = 2,fontsize = 2,risk.table.height = 1)
arrange_ggsurvplots(splots,print = TRUE,ncol = 2,nrow = 1)
```



One manager (Webster Butler) has worse long-term survival and remarkable drop at about 4 years of tenure. After manual inspection of the data, we can see that the parting employees didn't constitute one group, that would for instance be hired and left at the same period (eg. due to temporary increase of workload). Also, from his hiring date we notice that he was on the position for only one year and the departures were due earlier, meaning that the trend might be due to some other events in the company and not his management style. To analyse the managers' curves further, we see that nobody has such a big drop, so he might be managing the most difficult position. We also wanted to explore the flat line on the top of the graph: does this line consist of only well-payed individuals? This group is comprised of 19 people, assigned to 2 managers and a board of directors. Their survival time ranges between 1-11 years and they are of all age groups. Mostly they are well paid, besides for some exceptions.

```
par(mfrow=c(1,2))
barplot(table(none_df$AgeGrp),xlab = "Age group", main = "Age distribution in subgroup of employees")
barplot(table(none_df$PayGrp),xlab = "Pay group",main = "Salary distribution in subgroup of employees")
```

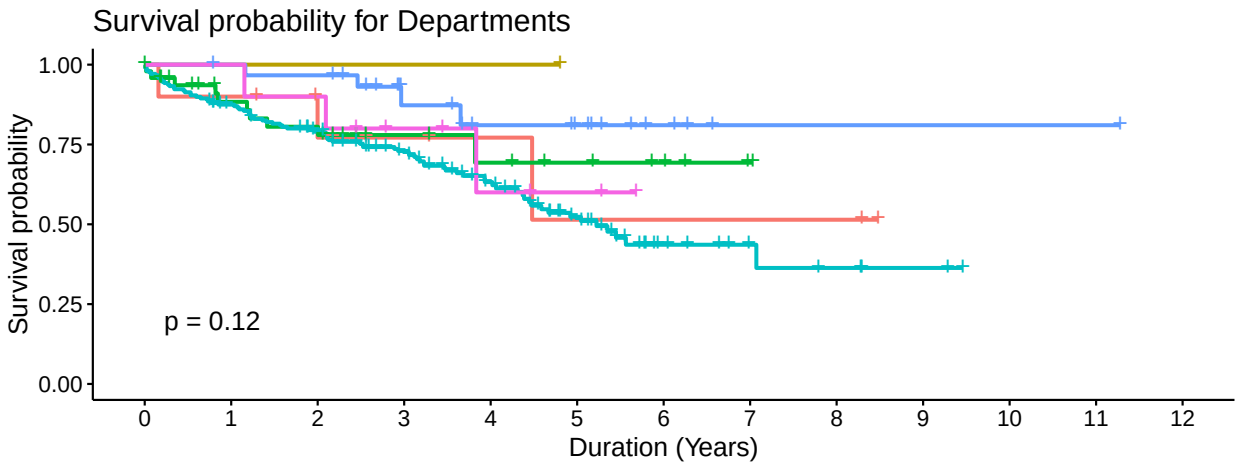




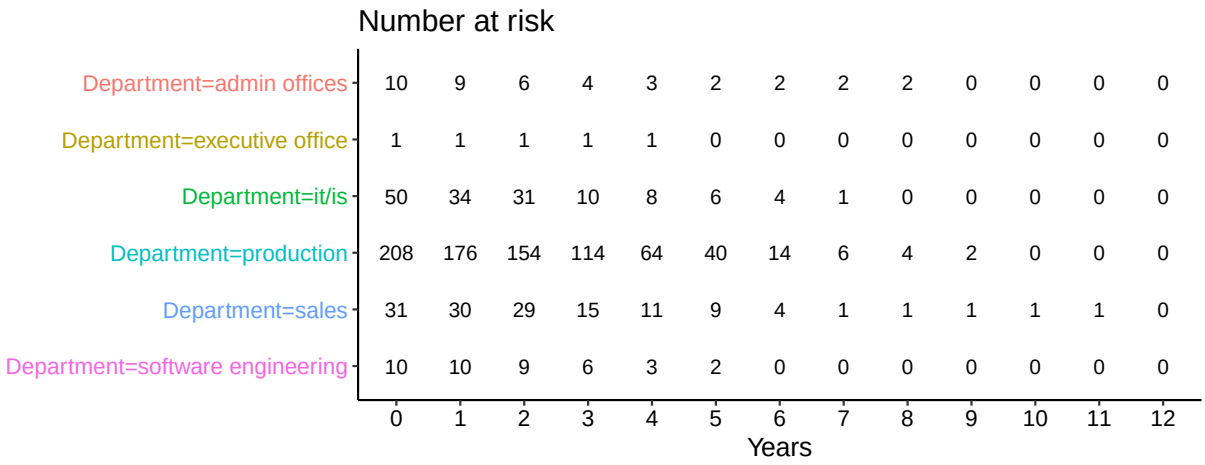
2.1.4 Department

Since we saw some difference in turnover between different managers, we verified also for different departments. The shape of the survival curve for Production department is quite different, but that might be due to much bigger group. We didn't find any significant difference, but at the same time, the p-value is not very high, so Department variable might still be used in more complex models.

```
dpt <- survfit(y ~ Department,data=df)
splots <- list()
splots[[1]]<- ggsurvplot(dpt,data=df,risk.table =F,censor=T,pval=T,xlab='Duration (Years)'
                         ,break.time.by = 1,legend='none',legend.title="",title = "Survival probability

arrange_ggsurvplots(splots,print = TRUE,ncol = 1,nrow = 1)
```




```
ggsurvplot(dpt,data=df, risk.table =T,censor=T, xlab='Years',break.time.by = 1,
legend='none',legend.title="",risk.table.fontsize = 4,fontsize = 2,risk.table.height = 1)
```



2.1.5 Summary of Kaplan-Meiers models

Manual check variable significant with KM-Models tested

Sex	PayGrp	Manager	Department
0.96	0.0088		0.00027

2.2 Cox proportionl hazard models

CPHM is semi-parametric regression model, which quantifies the hazard ratio between two groups (parametric part: beta, non-parametric: baseline hazard). In this chapter we would like to quantify the effect, observed by KM models earlier. In addition, we will construct different multiparametric models and compare their performances.

2.2.1 Salary level

Since Cox model is a regression model, we don't need to take grouped variable, but instead numeric value for PayRate.

```
payrate_cph<-coxph(y ~ I(PayRate/10), data=data)
#summary(payrate_cph)
```

We observe that for every 10 units of salary the chance of leaving drops for roughly 20%.

2.2.2 Manager

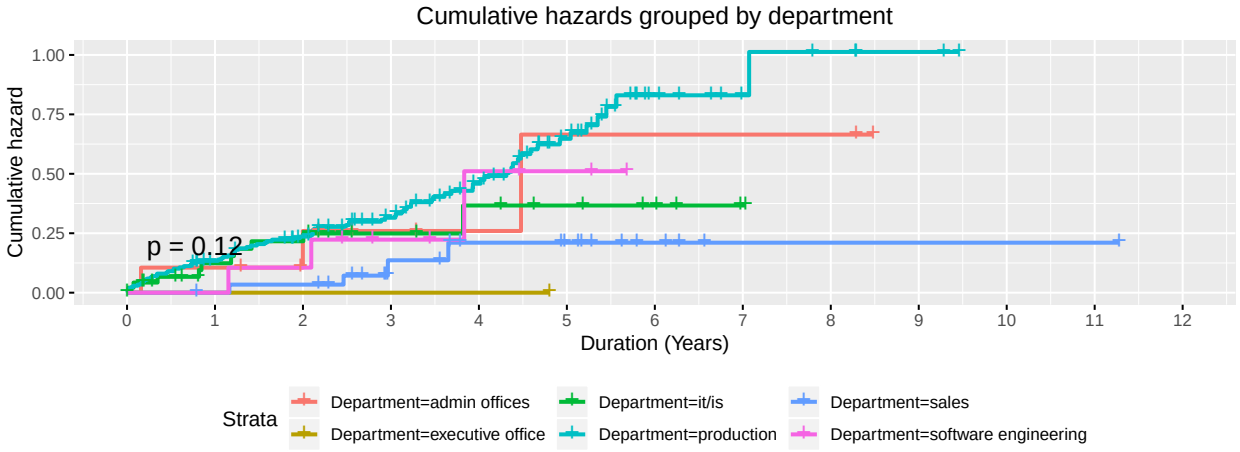
```
df<-mutate(df, MN = relevel(factor(MN), ref = "webster butler"))
mana_cph<-coxph(y ~ MN, data=df)
#summary(mana_cph)
```

We compared all the managers with Webster Butler. All managers have negative coefficients. 6 of them have p-value less than 0.05 and 3 more between 0.05 and 0.1.

Employees working for Webster Butler have about 3-times higher risk of leaving than the employees of managers having p-value less than 0.1.

2.2.3 Department

```
ggsurvplot(dpt, data=df, risk.table =F, censor=T,xlab='Duration (Years)',
           break.time.by = 1, legend='bottom',
           title="Cumulative hazards grouped by department",fun = "cumhaz",
           pval = T,
           ggtheme = theme(plot.title = element_text(hjust = 0.5)))
```



In the test, all of the groups were compared against the first group in alphabetical order ('admin offices'). In our case, the critical department according to KM was production. Thus, we set the production as a comparative group.

```
df<-mutate(df, Department = relevel(factor(Department), ref = "production"))
prod_cph<-coxph(y ~ Department, data=df)
summary(prod_cph)
```

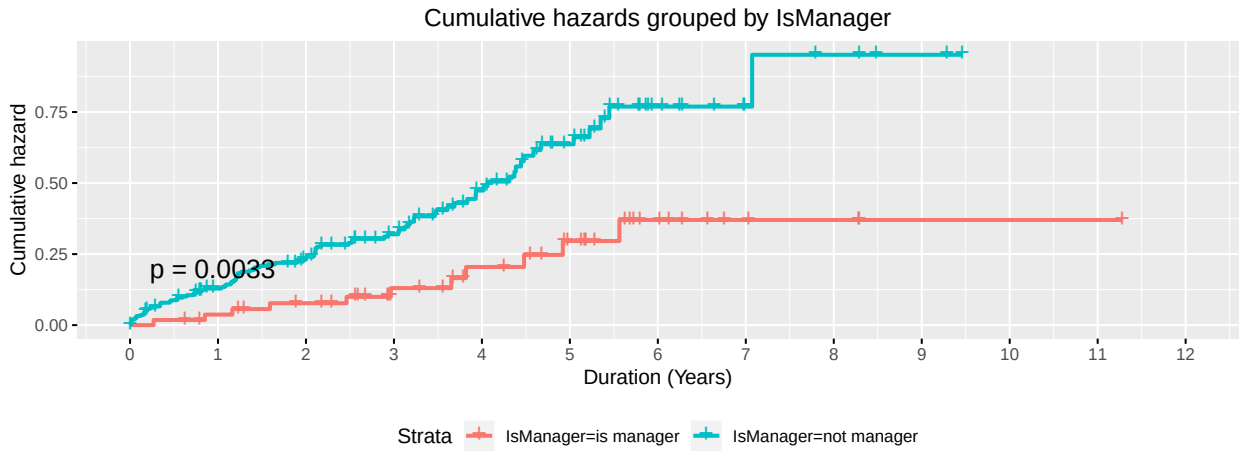
```
## Call:
## coxph(formula = y ~ Department, data = df)
##
##   n= 310, number of events= 103
##
##                                    coef  exp(coef)   se(coef)      z Pr(>|z|)
## Departmentadmin offices        -3.169e-01  7.284e-01  5.904e-01 -0.537   0.5914
## Departmentexecutive office     -1.520e+01  2.510e-07  2.507e+03 -0.006   0.9952
## Departmentit/is                -3.262e-01  7.216e-01  3.362e-01 -0.971   0.3318
## Departmentsales                -1.288e+00  2.759e-01  5.121e-01 -2.514   0.0119 *
## Departmentsoftware engineering -3.554e-01  7.009e-01  5.878e-01 -0.605   0.5455
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                exp(coef) exp(-coef) lower .95 upper .95
## Departmentadmin offices        7.284e-01  1.373e+00    0.2290    2.3171
## Departmentexecutive office     2.510e-07  3.984e+06    0.0000       Inf
## Departmentit/is                7.216e-01  1.386e+00    0.3734    1.3946
## Departmentsales                2.759e-01  3.624e+00    0.1011    0.7528
## Departmentsoftware engineering 7.009e-01  1.427e+00    0.2215    2.2184
##
## Concordance= 0.562  (se = 0.022 )
## Likelihood ratio test= 11.26  on 5 df,   p=0.05
## Wald test            = 7.31  on 5 df,   p=0.2
## Score (logrank) test = 8.74  on 5 df,   p=0.1
```

The only statistically significant difference is between production and sales (p=0.0119), with coefficient value -1.288 and exp(coef) = 0.28, meaning that employees in production will have around 3.6-times higher risk of leaving in comparison to sales department. Also, the coefficents for all other departments are negative, suggesting that employees have higher risk of leaving if they work in the production.

2.2.4 Managerial position

Is having a manager position influencing the duration?

```
mana <- survfit(y ~ IsManager,data=df)
ggsurvplot(mana, data=df, risk.table =F, censor=T,xlab='Duration (Years)',
           break.time.by = 1, legend='bottom',
           title="Cumulative hazards grouped by IsManager",fun = "cumhaz",
           pval = T,
           ggtheme = theme(plot.title = element_text(hjust = 0.5)))
```



If you're a manager, the risk of leaving is significantly higher. However the duration of stay is not affected.

2.3 Multivariate Cox regression analysis

2.3.1 Comparison of Cox models

Based on the following two articles (article 1 and article 2) we constructed 2 models. The first one is based on the performance at work : the important factors are absenteeism, disengagement and low productivity. In our dataset we have the following variables: performance score, engagement survey, employee satisfaction, special projects count in last 6 months and the number of times that the employee was late to work during the last 30 days. The latter being either NA or 0, we removed it from analysis. The second one is based on social criterias : Sex, Age, Marital status, Race, proximity recruitment.

```
M1 <- coxph(y ~ PerfScrGrp + EngagementSurvey + EmpSatGrp + SpecialProjectsCount, data =  data)
data<-mutate(data, RaceDesc = relevel(factor(RaceDesc), ref = "White"))
data<-mutate(data, MaritalDesc = relevel(factor(MaritalDesc), ref = "married"))
df<-mutate(df, MaritalDesc = relevel(factor(MaritalDesc), ref = "married"))
M2<-coxph(y ~ Sex + Age + MaritalDesc + RaceDesc + factor(ProximityRec), data =  data)
M3 <- coxph(y ~ I(PayRate/10) + MaritalDesc + PerfScrGrp + IsManager, data =df)
M4 <- coxph(y~ Sex+Department,data =df)
M5 <- coxph(y ~ Age + I(PayRate/10), data = df)
M6 <- coxph(y~ Sex + Age + I(PayRate/10), data = df)
M7 <- coxph(y ~ Sex + AgeGrp + PayGrp, data = df)
```

```
M8 <- coxph(y ~ Sex+ Age+ I(PayRate/10) + PerfScrGrp +EmpSatGrp, data = df)
M9 <- coxph(y ~ Age + I(PayRate/10) + EmpSatisfaction, data = df)
M10 <- coxph(y ~ I(PayRate/10) + PerfScrGrp, data=df)
```

```
summary(M1)
```

```
## Call:
## coxph(formula = y ~ PerfScrGrp + EngagementSurvey + EmpSatGrp +
##     SpecialProjectsCount, data = data)
##
##   n= 310, number of events= 103
##
##                          coef exp(coef) se(coef)      z Pr(>|z|)
## PerfScrGrpmedium     -0.59698   0.55047  0.33429 -1.786  0.07413 .
## PerfScrGrphigh       -1.31390   0.26877  0.47737 -2.752  0.00592 **
## EngagementSurvey     -0.03334   0.96721  0.07869 -0.424  0.67182
## EmpSatGrpmedium       0.66329   1.94117  0.58810  1.128  0.25938
## EmpSatGrphigh         0.86088   2.36524  0.58983  1.460  0.14442
## SpecialProjectsCount -0.04519   0.95582  0.05257 -0.860  0.39000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                      exp(coef) exp(-coef) lower .95 upper .95
## PerfScrGrpmedium        0.5505     1.8166    0.2859     1.060
## PerfScrGrphigh          0.2688     3.7207    0.1054     0.685
## EngagementSurvey        0.9672     1.0339    0.8290     1.128
## EmpSatGrpmedium         1.9412     0.5152    0.6130     6.147
## EmpSatGrphigh           2.3652     0.4228    0.7444     7.515
## SpecialProjectsCount    0.9558     1.0462    0.8622     1.060
##
## Concordance= 0.575  (se = 0.029 )
## Likelihood ratio test= 9.92  on 6 df,   p=0.1
## Wald test            = 9.49  on 6 df,   p=0.1
## Score (logrank) test = 9.79  on 6 df,   p=0.1
```

```
summary(M2)
```

```
## Call:
## coxph(formula = y ~ Sex + Age + MaritalDesc + RaceDesc + factor(ProximityRec),
##     data = data)
##
##   n= 310, number of events= 103
##
##                                              coef  exp(coef)   se(coef)      z Pr(>|z|)
## Sexm                                   -6.406e-02  9.379e-01  2.078e-01 -0.308   0.7579
## Age                                     1.390e-02  1.014e+00  1.105e-02  1.258   0.2084
## MaritalDescdivorced                     5.999e-01  1.822e+00  2.973e-01  2.018   0.0436 *
## MaritalDescseparated                   -1.695e+00  1.836e-01  1.016e+00 -1.668   0.0954 .
## MaritalDescsingle                      -4.005e-01  6.700e-01  2.267e-01 -1.766   0.0773 .
## MaritalDescwidowed                      7.169e-01  2.048e+00  5.342e-01  1.342   0.1796
## RaceDescAmerican Indian or Alaska Native -1.650e+01  6.832e-08  2.054e+03 -0.008   0.9936
## RaceDescAsian                          -6.610e-02  9.360e-01  3.294e-01 -0.201   0.8410
## RaceDescBlack or African American      -2.326e-01  7.925e-01  2.746e-01 -0.847   0.3970
## RaceDescHispanic                       -4.524e-01  6.361e-01  1.021e+00 -0.443   0.6577
## RaceDescTwo or more races              -2.408e-01  7.860e-01  4.153e-01 -0.580   0.5620
## factor(ProximityRec)1                   7.236e-02  1.075e+00  2.099e-01  0.345   0.7303
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                        exp(coef) exp(-coef) lower .95 upper .95
## Sexm                                   9.379e-01  1.066e+00   0.62411     1.410
```

```
## Age                                      1.014e+00  9.862e-01   0.99227     1.036
## MaritalDescdivorced                      1.822e+00  5.489e-01   1.01726     3.263
## MaritalDescseparated                     1.836e-01  5.447e+00   0.02504     1.346
## MaritalDescsingle                        6.700e-01  1.493e+00   0.42960     1.045
## MaritalDescwidowed                       2.048e+00  4.883e-01   0.71879     5.836
## RaceDescAmerican Indian or Alaska Native 6.832e-08  1.464e+07   0.00000       Inf
## RaceDescAsian                            9.360e-01  1.068e+00   0.49076     1.785
## RaceDescBlack or African American        7.925e-01  1.262e+00   0.46266     1.357
## RaceDescHispanic                         6.361e-01  1.572e+00   0.08598     4.706
## RaceDescTwo or more races                7.860e-01  1.272e+00   0.34830     1.774
## factor(ProximityRec)1                    1.075e+00  9.302e-01   0.71240     1.622
##
## Concordance= 0.637  (se = 0.029 )
## Likelihood ratio test= 24.34  on 12 df,   p=0.02
## Wald test            = 19.92  on 12 df,   p=0.07
## Score (logrank) test = 23  on 12 df,   p=0.03
```

The only important variable from the first model seems to be the Performance Score Group. The only significant variable from the second model is marital status (we were comparing with married employees). Interestingly, divorced have positive coefficient, while separated and single have a negative one.

We implemented additional models to determine what covariates would influence an employee from quitting the company. The best model was based on the following covariates: Sex, Age, PayRate, PerfScrGrp and EmpSatGrp.

```
fits <- list(M1 = M1, M2=M2, M3=M3, M4=M4, M5=M5, M6=M6, M7=M7, M8=M8, M9=M9, M10 = M10)
sapply(fits, AIC)
```

```
##       M1       M2       M3       M4       M5       M6       M7       M8       M9      M10
## 1067.775 1065.357 1049.500 1066.412 1059.441 1061.383 1061.536 1060.214 1064.722 1057.358
```

Here the best model is M3 (PayRate, MaritalDesc, PerfScrGrp and IsManager) by AIC.

```
summary(M3)
```

```
## Call:
## coxph(formula = y ~ I(PayRate/10) + MaritalDesc + PerfScrGrp +
##     IsManager, data = df)
##
##   n= 310, number of events= 103
##
##                           coef exp(coef) se(coef)      z Pr(>|z|)
## I(PayRate/10)         -0.05421   0.94723  0.10145 -0.534   0.5931
## MaritalDescdivorced    0.58368   1.79262  0.29897  1.952   0.0509 .
## MaritalDescseparated  -1.57476   0.20706  1.01178 -1.556   0.1196
## MaritalDescsingle     -0.38139   0.68291  0.22346 -1.707   0.0879 .
## MaritalDescwidowed     0.78530   2.19307  0.52770  1.488   0.1367
## PerfScrGrpmedium      -0.44628   0.64000  0.29354 -1.520   0.1284
## PerfScrGrphigh        -1.15518   0.31500  0.45233 -2.554   0.0107 *
## IsManagernot manager   0.77600   2.17277  0.43967  1.765   0.0776 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                       exp(coef) exp(-coef) lower .95 upper .95
## I(PayRate/10)            0.9472     1.0557    0.7764    1.1556
## MaritalDescdivorced      1.7926     0.5578    0.9977    3.2209
## MaritalDescseparated     0.2071     4.8296    0.0285    1.5042
## MaritalDescsingle        0.6829     1.4643    0.4407    1.0582
## MaritalDescwidowed       2.1931     0.4560    0.7796    6.1692
## PerfScrGrpmedium         0.6400     1.5625    0.3600    1.1377
```

```
## PerfScrGrphigh          0.3150     3.1746    0.1298    0.7644
## IsManagernot manager    2.1728     0.4602    0.9178    5.1435
##
## Concordance= 0.676  (se = 0.027 )
## Likelihood ratio test= 32.2  on 8 df,   p=9e-05
## Wald test            = 27.09  on 8 df,   p=7e-04
## Score (logrank) test = 29  on 8 df,   p=3e-04
```

For each PayRate/10, there is a ~6% less chance of leaving the company while before it was 20% (when doing Cox regression with only PayRate/10).

```
res.cox1 <- M3
test.ph1 <- cox.zph(res.cox1, transform = "km", global=TRUE)
test.ph1
```

```
##                            rho  chisq      p
## I(PayRate/10)        -0.1001  1.152 0.2832
## MaritalDescdivorced  -0.1588  2.669 0.1023
## MaritalDescseparated  0.1721  3.094 0.0786
## MaritalDescsingle     0.0634  0.416 0.5189
## MaritalDescwidowed   -0.0394  0.160 0.6887
## PerfScrGrpmedium     -0.0494  0.252 0.6155
## PerfScrGrphigh        0.1342  1.838 0.1752
## IsManagernot manager -0.1384  2.527 0.1119
## GLOBAL                   NA 13.239 0.1039
```

Considering that p-values are not small, it means that there are no time dependent coefficients. Which means that the hazard rate of groups in tested covariates is relatively constant in time.

3. Machine learning approach

In order to find out which factors would affect the risk of leaving, we built different models following a machine learning approach. In this section, we used some simple machine learning method such as :

- Elastic-Net Regularized Generalized Linear Models (`glmnet`)
- Coxph models selected by aic step function (`step`)
- Parameters chosen from the section above.

In order to analyse, we first have to build another dataset for the modeling part. We have to :

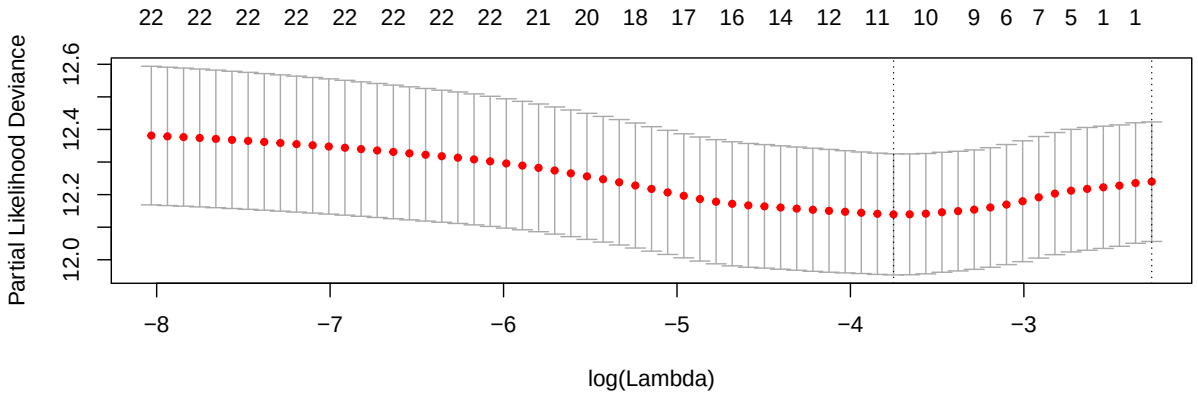
- Change the binary categorical variables to zero and one
- Remove some duplicated information (such as keep the Age and remove AgeGrp)
- Rescale some small value variables to have bigger effect on the models

```
mdf<-mutate(df,IsManager=ifelse(IsManager=="is manager",1,0)#1 is manager
            ,Sex=ifelse(Sex=='m ',1,0)#1 is male
            ,EngagementSurvey_10=I(EngagementSurvey/10)
            ,SpecialProjects_10=I(SpecialProjectsCount /10)
            ,PayRate_10=I(PayRate/10))

mdf<-subset(mdf,select=-c(MN,Position,AgeGrp,EmpSatisfaction,PayGrp,
                  EngagementSurvey,RecruitmentSource,SpecialProjectsCount))

mx<-subset(mdf,select = -c(Termd,Duration))
my<-with(mdf,Surv(Duration ,Termd))
#summary(mdf)
```

3.1 Linear model: Elastic net

In the elastic net model the regularization path is computed for the elasticnet penalty at a grid of values for the regularization parameter λ. We selected the best coefficent using λ that gives minimum mean cross-validated error to select the variables that provide the most regularized model.

```
set.seed(123)
X<-model.matrix(my~.,data=mx)
Mglmnet <-cv.glmnet(X,my,family="cox")
plot(Mglmnet)
```




```
b.enet.all<-coef(Mglmnet,s="lambda.min")
suppressMessages(b.enet<-b.enet.all[b.enet.all != 0])
names(b.enet) <- colnames(X)[as.logical(b.enet.all != 0)]
```

The plot of cross-validated partial log-likelihood deviance, including upper and lower standard deviations, as a function of logλ for the data set. The dotted vertical lines indicate the λ values with minimal deviance (left) and with the largest λ value within one standard deviation of the minimal deviance (right).

As one standard deviation of the minimal deviance didn't select any variables, here we chose to use minimal deviance λ with 11 variables selected as shown below:

```
##                          Age        MaritalDescdivorced       MaritalDescseparated
##                  0.008417999                0.381105409               -0.768840593
##           MaritalDescsingle         MaritalDescwidowed                  IsManager
##                -0.262253504                0.399934068               -0.538800225
## Departmentexecutive office            Departmentsales             PerfScrGrphigh
##                -0.605499215               -0.069556026               -0.458413214
##               EmpSatGrphigh                 Production
##                  0.022867471                0.147776742
```

3.2 Cox models

3.2.1 Manual and AIC models

- Manual model

To build the manual model we used the information from the above analysis. We deduced some important factors: payrate, manager, performance score and marital status. The variable Department doesn't appear

to have an obvious impact on the model. Therefore we used the nested models test `anova` to check if adding in Department variable would have an influence. The null hypothesis is that there's no difference between the 2 nested models. The result show that the p-value is high and that there's no difference. We can therefore take out the variable Department for future models.

```
Mmanual<-coxph(my~ PayRate_10+MaritalDesc+IsManager+PerfScrGrp,data=mx)
Mmanual1<-coxph(my~ PayRate_10+MaritalDesc+IsManager+PerfScrGrp+Department,data=mx)
anova(Mmanual,Mmanual1)
```

```
## Analysis of Deviance Table
##  Cox model: response is  my
##  Model 1: ~ PayRate_10 + MaritalDesc + IsManager + PerfScrGrp
##  Model 2: ~ PayRate_10 + MaritalDesc + IsManager + PerfScrGrp + Department
##    loglik  Chisq Df P(>|Chi|)
## 1 -516.75
## 2 -515.40 2.7036  5    0.7456
```

```
b.manual<-coef(Mmanual)
```

- AIC Step-model

In the AIC model we used the `step` functions with backward method to choose the best aic score of different combinations of variables for the Cox model.

```
Maic<-step(coxph(my~.,data=mx),trace = FALSE)  #coxph(my~.,data=mx) is the full model
b.aic<-coef(Maic)
Maic
```

```
## Call:
## coxph(formula = my ~ Age + MaritalDesc + IsManager + PerfScrGrp,
##     data = mx)
##
##                           coef exp(coef) se(coef)      z       p
## Age                    0.01860   1.01877  0.01105  1.682 0.09247
## MaritalDescdivorced    0.58594   1.79667  0.29579  1.981 0.04760
## MaritalDescseparated -1.68637   0.18519  1.01469 -1.662 0.09652
## MaritalDescsingle    -0.36564   0.69375  0.22322 -1.638 0.10142
## MaritalDescwidowed    0.81333   2.25540  0.52772  1.541 0.12326
## IsManager            -0.96666   0.38035  0.32414 -2.982 0.00286
## PerfScrGrpmedium     -0.49648   0.60867  0.29371 -1.690 0.09096
## PerfScrGrphigh       -1.20308   0.30027  0.45201 -2.662 0.00778
##
## Likelihood ratio test=34.61  on 8 df, p=3.151e-05
## n= 310, number of events= 103
```

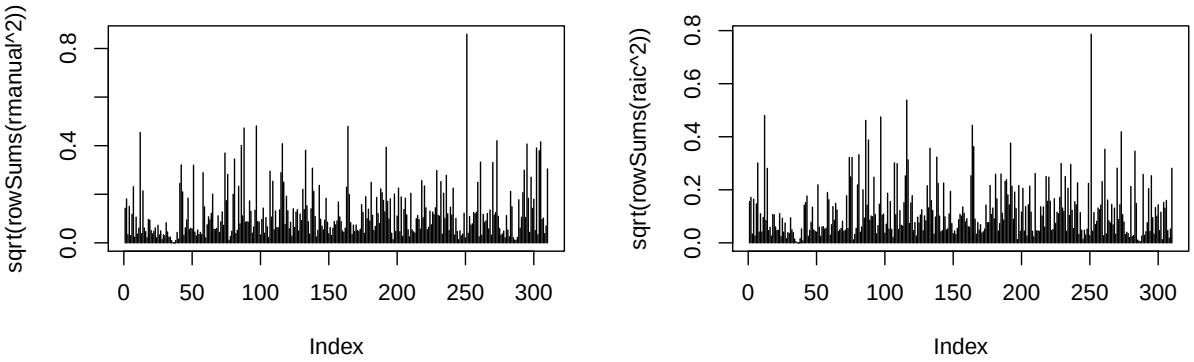
3.2.2 Cox model diagnostics

For Cox models (manual and aic), we verified if there were outliers by checking the residuals of the models. We also checked the proportional hazards assumption for Cox models.

- Cox model residuals

We plot the residuals, with `type = 'dfbetas'`. From plots below, we didn't see extreme outliers.

```
rmanual<-residuals(Mmanual,type = 'dfbetas')
raic<-residuals(Maic,type = 'dfbetas')
par(mfrow = c(1,2))
plot(sqrt(rowSums(rmanual^2)),type='h')
plot(sqrt(rowSums(raic^2)),type='h')
```



- Proportionality of risk

Schoenfeld residuals are used to verify the proportionality of the risk. Using null hypothesis being that the model is proportional. We do not reject the null hypothesis (p-value bigger than 0.05). We can see that for all variables as for the global p-value in both models (Mmanual and Maic) , we do not reject the proportionality.

```
cox.zph(Mmanual)
```

```
##                          rho  chisq      p
## PayRate_10           -0.1001  1.152 0.2832
## MaritalDescdivorced  -0.1588  2.669 0.1023
## MaritalDescseparated  0.1721  3.094 0.0786
## MaritalDescsingle     0.0634  0.416 0.5189
## MaritalDescwidowed   -0.0394  0.160 0.6887
## IsManager             0.1384  2.527 0.1119
## PerfScrGrpmedium     -0.0494  0.252 0.6155
## PerfScrGrphigh        0.1342  1.838 0.1752
## GLOBAL                    NA 13.239 0.1039
```

```
cox.zph(Maic)
```

```
##                          rho   chisq      p
## Age                   0.1578  2.8838 0.0895
## MaritalDescdivorced  -0.1547  2.5753 0.1085
## MaritalDescseparated  0.1574  2.4892 0.1146
## MaritalDescsingle     0.0693  0.4918 0.4831
## MaritalDescwidowed   -0.0259  0.0698 0.7916
## IsManager             0.1073  1.2024 0.2728
## PerfScrGrpmedium     -0.0753  0.5813 0.4458
## PerfScrGrphigh        0.1166  1.3772 0.2406
## GLOBAL                    NA 14.7908 0.0633
```

- Stratification

However, since the p-value of the AIC model is not far from 0.05, we can try to stratify some variables to see if the model gets better (using the concordance of the model). We stratified the variables that has p-value less then 0.1 on the above 2 models : MaritalDesc and Age. The results showed that the concordance of the new models didn't improve. We won't stratify values in the following models.

```
N_Mmanual<-coxph(my~ PayRate_10+strata(MaritalDesc)+IsManager+PerfScrGrp,data=mx)
N_Maic<-coxph(my~strata(Age)+MaritalDesc+IsManager + PerfScrGrp,data=mx)

df_global_p<-tibble(model = c("manual","aic"),
                    origin_conc=c(summary(Mmanual)$concordance[1],summary(Maic)$concordance[1]),
                    new_conc=c(summary(N_Mmanual)$concordance[1],summary(N_Maic)$concordance[1]))
df_global_p
```

```
## # A tibble: 2 x 3
##   model  origin_conc new_conc
##   <chr>        <dbl>    <dbl>
## 1 manual       0.676    0.606
## 2 aic          0.679    0.669
```

3.3 Models variable summary

Below we can see a summary of all the chosen variable among all three candidate models. We see that MaritalDesc, IsManager and PerfScrGrp are selected in the models.

```
names(b.enet)
```

```
##  [1] "Age"                          "MaritalDescdivorced"         "MaritalDescseparated"
##  [4] "MaritalDescsingle"            "MaritalDescwidowed"          "IsManager"
##  [7] "Departmentexecutive office" "Departmentsales"             "PerfScrGrphigh"
## [10] "EmpSatGrphigh"                "Production"
```

```
Mmanual$formula
```

```
## my ~ PayRate_10 + MaritalDesc + IsManager + PerfScrGrp
```

```
Maic$formula
```

```
## my ~ Age + MaritalDesc + IsManager + PerfScrGrp
```

4. Model selection

We have three candidate models above. In this section we compare their performance using the models' coefficients in order to elect the best one.

```
#We get all the coefficients of models
models_coefficients <- tibble(model = c("enet","manual","aic"),
                              coefficients = list(b.enet,b.manual,b.aic))
```

4.1 Log rank ratio

First method for performance: log hazard ratio. We use the coefficient retrieved from above models to build the prediction in order to construct the Log rank ratio for all models. The estimates (the bigger the better) and p-values show that the AIC model performs the best.

```
lincom <- function(b,X) rowSums(sweep(X[,names(b),drop = FALSE],2,b,FUN = "*"))
X.new=model.matrix(my~.,data=mx)
models_performance <- mutate(models_coefficients,
                      predictions = map(coefficients,~ lincom(.,X.new)),
                      cox_obj = map(predictions,~ coxph(my ~ I(. / sd(.)))),
                      cox_tab = map(cox_obj,broom::tidy)) %>% unnest(cox_tab)
models_performance[c(1,6,7,8)]
```

```
## # A tibble: 3 x 4
##   model  estimate std.error statistic
##   <chr>     <dbl>     <dbl>     <dbl>
## 1 enet      0.640     0.118      5.44
## 2 manual    0.637     0.122      5.20
## 3 aic       0.656     0.121      5.44
```

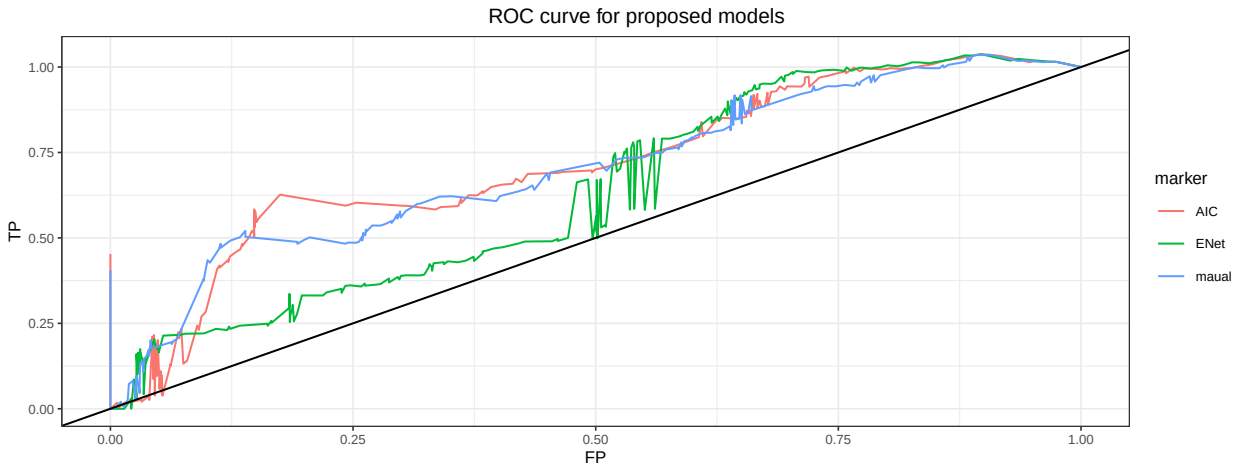
4.2 AUC + ROC plot

Another value we can check is the ROC curve, to check different thresholds affecting false positive and true positive rate. We would prefer to have higher AUC rate. We can see from the numeric value and the plot that here AIC still is the best model. The prediction time we used is 11.5 years due to the max duration of the whole dataset is around 11.5 years. We are comparing the prediction with the true value of the entire dataset.

```
ROC.enet=survivalROC(my[,1],my[,2],models_performance$predictions[[1]],
                     predict.time = 365.25 * 11.5,method = "KM")
ROC.manual=survivalROC(my[,1],my[,2],models_performance$predictions[[2]],
                       predict.time = 365.25 * 11.5,method = "KM")
ROC.aic=survivalROC(my[,1],my[,2],models_performance$predictions[[3]],
                    predict.time = 365.25 * 11.5,method = "KM")
ROC<-list(ENet=ROC.enet,maual=ROC.manual,AIC=ROC.aic)
map_dbl(ROC,"AUC")
```

```
##      ENet     maual       AIC
## 0.6340869 0.7161101 0.7271632
```

```
df<- map(ROC,~ with(.,tibble(cutoff = cut.values,FP,TP)))
for(nm in names(df)) {df[[ nm ]]$marker <- nm}
dat <- do.call(rbind,df)
ggplot(dat,aes(FP,TP,color = marker)) +
  geom_line() +
  geom_abline(slope = 1)+
  ggtitle("                                                                    ROC curve for prop
  theme_bw(base_size = 9)
```



4.3 Final model selection

We summarize here using the value of each performance score and give it a weighted score (the score divided by the highest score , hence 100 is the best). Note that we don't have concordance score for enet model as well as the concordance for AIC and manual model are very similar, we didn't take it for the comparison.

```
## # A tibble: 3 x 6
##   model    LRR LRR_Score   ROC ROC_Score AVG_Score
##   <chr>  <dbl>     <dbl> <dbl>     <dbl>     <dbl>
## 1 enet   0.64         98 0.634        87        92
## 2 manual 0.637        97 0.716        98        98
## 3 aic    0.656       100 0.727       100       100
```

After comparing 2 different scores the result for the final model selection is clearly AIC model. It is also interesting to see that the manual model, of which the variables were chosen by our own anaysis, perform very well and almost similar to AIC function. There is only one variable different in both models, three other variables are the same (MaritalDesc, IsManager and PerfScrGr).

5. Summary

The purpose of the project was to determine which factors contribute the most to the turnover of the employees. We used the following methodologies: Kaplan-Meier, Cox models, Machine Learning (ElasticNet, AIC), with different selections of variables. We determined that the most important are : Marital status, IsManager, Performance Score and Age. However depending on the method, the PayRate and Manager showed to be significant.