Data Mining Project

Daily Happiness & Employee Turnover DataSet

Hello,

In this report, I will explain what I did to analyse the Dataset I choose by respecting the following six steps designed for this :

Problem Understanding:

My present project will deal with the relationship between employee happiness and satisfaction in companies and employee turnover, so the question is: can happiness predict the turnover of employees in companies? or is it the opposite?

The dataset is made to answer this question mainly, to give you a little idea about its creation and author: these data were exploited by Jose Berengueres during the summer of 2016 in Barcelona when he attended a conference where Alex Rios - the managing director of myhappyforce.com - explained one of his products. He built an app where employees report daily levels of happiness at work. This application is used by companies to track the happiness of workers. After the conference, Jose asks him if he wants to use the sources to better understand the phenomenon of staff turnover, and that's how the idea of analysing these data took place.

So, to analyse this dataset, I will try to get some inspiration from the treatments already done on the kaggle website, to make things a little easier and to try to better understand if there is a relationship between employee satisfaction, expressed by their votes and their turnover.

Inspiration: (from the description of the dataset)

The cost of employee turnover has been pointed out extensively in the literature. A high turnover rate not only increases human resource costs, which can reach up to 150% of the annual salary per replaced employee, but it also has social costs, as it is correlated with lower wages, lower productivity per employee, and not surprisingly, a less loyal workforce.

For reference, in 2006, turnover at Walmart's Sam's Club was 44% with an average hourly pay of $10.11, while at Costco it was a much lower 17% with a higher $17.0 hourly wage.

In addition, a more recent study correlated companies with low turnover with a series of socially positive characteristics dubbed high-involvement work practices.

On the other hand, research on employee turnover (churn) is not a prolific topic in the engineering community. In IEEE publications, one can find just over 278 publications with titles containing the keyword churn, and the bulk of those focus on customer churn, and specifically churn in the telecommunications industry, while on the topic of employee churn there is just one title indexed.

The goal is to clarify the characteristics of employees that will churn (or that are at risk of churning), to help companies understand the causes so they can reduce the turnover rate.

Data Understanding:

The data includes four tables: votes, comments, interactions and churn.

Vote table:

A vote was obtained when an employee opened the application and answered the question: How happy are you at work today? To vote, the employees indicate their feelings by touching one of the four icons on the screen (corresponding to 1 to 4 as notation).

Comments table:

Once the employee has indicated his level of happiness, a second screen appears in which he can enter a textual explanation (usually a complaint, a suggestion or a comment). This is the comment table. Out of 4,356 employees, 2,638 employees commented at least once.

Interactions table:

Finally, in a third screen, the employee can see the comments of his peers and like them or not, these data are stored in the table of interactions. 3,516 employees liked or did not like at least one of their peers' comments.

Churn table:

The churn table contains the time when an employee resigned (left or was fired).

Data Preparation:

In this step, I loaded the dataset into R with its 4 tables, and I tried to clean them, and eventually delete the unimportant columns.

In this step, I'm inspired by what Jose did, because this part is very important for the next steps, but I added some manipulations that he did not do, such as cleaning NA entries for example.

Interactions table:

I delete this entry that does not correspond, with the instruction:

```
likes = likes[likes$companyAlias!="58a728a",]
```

After that, I transformed the variable companyAlias from an id without meaning to a simple alias with few letters:

```
likes = droplevels(likes)
levels(likes$companyAlias) =
c("A","B","C","D","E","F","G","H","I","J","K","L","M","N","O","P","Q","R","S","T","V","W","X","Y","Z","AA
","AB","AC","AD","AE","AF","AG","AH","AI","AJ")
likes$companyAlias = as.factor(likes$companyAlias)
```

And finally I have attribute to each entry an unique uid by:

```
likes$uid = paste(likes$employee,likes$companyAlias)
```

Comments table:

I removed duplicate elements/rows by: **comments = unique(comments)**

I removed those companies that has low numbers and no comments:

```
comments <- comments[comments$coa!="5474b9cde4b0bf7614b2c66f",]
comments <- comments[comments$coa!="58bf03e5cff4fa0004dd44ef",]
```

I did the same thing for companyAlias as in the "likes" table to have a simple alias:

After that I parsed dates to have an appropriate format like:

"2017-03-20" instead of "Thu Feb 23 12:48:04 CET 2017".

And finally I make a unique id for all employees by :

```
comments$uid <- paste(comments$id,comments$coa)
```

Votes table:

I removed those companies that has low numbers and no comments:

```
votes<-votes[votes$companyAlias!="5474b9cde4b0bf7614b2c66f",]
votes<-votes[votes$companyAlias!="58bf03e5cff4fa0004dd44ef",]
votes<-votes[votes$companyAlias!="573a0671b5ec330003add34a",]
```

I did the same thing for companyAlias as in the "likes" table to have a simple alias.

After that I parsed dates to have an appropriate format like : "2017-03-20" instead of "Thu Feb 23 12:48:04 CET 2017".

And a unique id as for the two last tables by :

```
votes$uid <- paste(votes$employee,votes$companyAlias)
```

Churn table:

The same thing as for the previous tables, I removed the companies that has low numbers and no comments, I changed the companyAlias as in the "likes" table to have a simple one, after that I parsed dates to have an appropriate format like : "2017-03-20" instead of "Thu Feb 23 12:48:04 CET 2017", after that I changed the churn values from True False to 1 and 0, after that I removed duplicated and negative id users from churn.

Finally for this step, I have removed all NA entries by:

```
likes = na.omit(likes);          comments = na.omit(comments);
votes = na.omit(votes);          churn = na.omit(churn)
```

Modeling and Evaluation:

In this two parts of the analysis, I will do some manipulations on the Dataset and each time I will give some explanations to what I obtain (**Evaluations, that's why this two parts are merged together**) (this explanations can be false some times of course ..., but it reflect my understanding for the data).

The first thing I have done is some descriptive statistics:

How much happiness there is by company:

```
happinessMean <- aggregate(vote ~ companyAlias, votes, function(x) mean(x))
colnames(happinessMean)<- c("Company","MeanHappyIndex")
```

The same result but this time ordered by decreasing mean:

```
happinessMean <- happinessMean[with(happinessMean, order(-MeanHappyIndex, Company)), ]
```

The mean rate are from 2.48 to 3.5.

What is the total mean of votes : **mean(votes[,4])** # I got 2.86, and that's a good result, which I did not expect because nobody likes his job, according to some studies.

sd(votes[,4]) # the standard deviation was 0.98

The distribution of notations : **barplot(table(votes[,4]))**

Most employee have 3 as vote rate, the second position is for 4 and after that 1 and 2

```
df = data.frame(table(votes[,4]))
df[,2] = df[,2] / sum(df[,2]) * 100
bp=ggplot(df, aes(x="Votes", y=Freq, fill=Var1))+geom_bar(width = 1, stat = "identity")
pie <- bp + coord_polar("y", start=0)
```

pie # almost the half of notations has 3 as value, the same result but with a plot.

Some histograms that describe the data:

active users per day: **ggplot(votes[votes$voteDate>max(votes$voteDate)-380,], aes(voteDate,fill=companyAlias))+geom_bar(alpha = .95) + labs(title = "Active users per day")**

This plot shows that the number of active users increase through time. For some companies the number of active users was 0 before Oct 2016 like AA to AI.

Distribution of happiness seen differently by company:

```
ggplot(votes, aes(vote,fill=companyAlias)) +  geom_bar(alpha = 1) + labs(title = "Distribution of happiness")
```

This plot shows that we have the same result as before but this time with companies' alias that show which company is best rated.

Distribution of happiness by day of week and by company:

```
votes$weekday <- weekdays(votes$voteDate)
ggplot(votes[votes$voteDate>max(votes$voteDate)-380,], aes(weekday,fill=companyAlias))+
geom_bar(alpha = .95)
```

We can see here with this plot that there is something weird, the number of votes in the weekend (Saturday and Sunday) is very low compared to the rest days of the week, maybe because the weekend is for rest or when they are at work employees encounter situations that remind them the vote...

Now I will check the happiness / weekday distribution:

```
hbyw <- aggregate(vote ~ weekday, votes, function(x) mean(x))
```

```
colnames(hbyw) <- c("weekday","MeanHappyIndex")
hbyw <- hbyw[with(hbyw, order(-MeanHappyIndex)), ]
qplot(weekday, data=hbyw, geom="bar", weight= MeanHappyIndex, ylab="Happyness")
wee<- rbind(rep(3.0,7) , rep(2.7,7), hbyw$MeanHappyIndex)
colnames(wee)<- hbyw$weekday
radarchart(as.data.frame(wee),axistype=1,cglcol="grey", cglty=1, axislabcol="black",
caxislabels=seq(2.7,3.0,0.05), cglwd=0.8,pcol=rgb(0.2,0.5,0.5,0.6) , pfcol=rgb(0.2,0.5,0.5,0.5) , plwd=4 ,
vlcex=1 ,title="A Tuesday is the least happy day ")
```

This plot show that Tuesday is the least happy day, the day that have the lower rates.

Influence of weekday on happiness:

```
hbye <- aggregate(vote ~ uid+weekday, votes, function(x) mean(x))
colnames(hbye)<- c("uid","weekday","MeanHappyIndex")
ggplot(hbye, aes(MeanHappyIndex,fill=weekday)) +  geom_bar(alpha = 0.5)+ geom_histogram(binwidth = .1)
+ labs(title = "Influence of weekday on happiness")
```

Here we can see the last result but with more details on each rate.

Average vote by company :

```
v.co.s <- aggregate(votes$vote, by=list(Category=votes$companyAlias), FUN=sum)
v.co.m <- aggregate(votes$vote, by=list(Category=votes$companyAlias), FUN=mean)
v.co.sd <- aggregate(votes$vote, by=list(Category=votes$companyAlias), FUN=sd)
vote.by.co <- cbind(v.co.s,v.co.m$x,v.co.sd$x)
colnames(vote.by.co) <- c("alias","sum","mean","sd")
vote.by.co
```

The same thing as before but this time we have the mean and sd for all companies.

The distribution of churns in company through time:

```
ggplot(churn[churn$stillExists==0,], aes(lastParticipationDate+15,fill=companyAlias))
+geom_histogram(alpha = .95,binwidth=7) + xlab("Churn date")
```

This plot allows us to see that churn increase through time between 2015 and 2017.

Finally to understand the relation between employees satisfaction (happiness) and churn (turnover) I will plot the mean of votes for each user with respect to churn:

```
satisfaction = aggregate(vote ~ employee, votes, function(x) mean(x))
```

For each user we have now his vote mean in this list or table. I will plot this result:

```
plot(satisfaction$employee, satisfaction$vote)
```

Now I will compute how many times each user has been churned: 0: never, 1: one time, 2: two times....

```
churned = aggregate(churn ~ employee, churn, function(x) sum(x))
```

We have now for each user the number of churn in this table churned.

It's time now to see with color the distribution of churn (turnover) with respect to satisfaction (happiness): I will first combine the two tables in first one “satisfaction”:

satisfaction$churn = churned$churn, we have now a third column in this table that contains the churn information of employees.

```
plot(satisfaction[satisfaction$churn==0,]$employee, satisfaction[satisfaction$churn== 0,]$vote, pch = 20, col = "green",main="Average employee vote / the number of times they were churned", xlab = "employees", ylab = "mean of votes")
abline(h=2.75, col="black")  # I add a line that split the points at the mean vote = 2.5.
```

We see clearly that most people who have never been fired have an average vote >= 2.75, It can mean that they are satisfied by their jobs. I will add the employees who have a churn value == 1 with different color and I will put a legend for the plot:

```
points(satisfaction[satisfaction$churn==1,]$employee, satisfaction[satisfaction$churn==1,]$vote, pch = 20, col = "red")
legend(x=0,y=1.75,c("0 times", "1 time"), cex=.8,col=c("green", "red"),pch=c(20))
```

Also for employees hired once, we can see that they have a good average vote that is >= 2.75, the same conclusion can be valid for them also. For the rest of the employees who were churned at least 2 times, we can see that in most cases, the average is not very important even if it's not perfect because of course we'll have other factors that influence the turnover in addition to happiness like money, family ...

So I can say at this point that it's an openness to my analysis of this dataset, analyse and observe the influence of other factors of employee turnover in companies like money and social status, seems a very important subject to study.

in the folowing you can see the distribution of other employees who were fired 2 or more times :

```
points(satisfaction[satisfaction$churn==2,]$employee, satisfaction[satisfaction$churn==2,]$vote, pch = 20, col = "cyan")
points(satisfaction[satisfaction$churn==3,]$employee, satisfaction[satisfaction$churn==3,]$vote, pch = 20, col = "purple")
points(satisfaction[satisfaction$churn==4,]$employee, satisfaction[satisfaction$churn == 4,]$vote, pch = 20, col = "blue")
```

This two last cases are not very important because we have just one employee per case : we can consider them as outlayers :

```
points(satisfaction[satisfaction$churn==5,]$employee, satisfaction[satisfaction$churn == 5,]$vote, pch = 20, col = "black")
points(satisfaction[satisfaction$churn==6,]$employee, satisfaction[satisfaction$churn == 6,]$vote, pch = 20, col = "black")
```

References:

The only two references I used in my work are:

- The kaggle website, where I found my Dataset, some parts of code are taken from this source:

  https://www.kaggle.com/harriken/employeeturnover

- The course of Mr. Fabrice Muhlenbach of Data Analysis 1st semester.