Students
Aleksandra Budkina
Jason Chau
Woo Lee
Wenyue Li
Ke Miao

Introduction.

The group's topic of interest was to analyze data collected from the employees which possibly can be a measurement for employees' performance, involvement, and satisfaction with the current position and answer a question, whether or not that employee left. To do so we decided to use a simulated dataset from kaggle: "Human Resources Analytics. Why are our best and most experienced employees leaving prematurely?" from <https://www.kaggle.com/ludobenistant/hr-analytics>.

I. Description of data, explanatory variables, response variable.

The total number of entries in our dataset is 14999. None of the entries are missing.
The dataset contains the following variables:

- satisfaction_level (how satisfied was the employee)
- last_evaluation (how did the company last evaluate the employee)
- number_project (number of projects)
- average_montly_hours (average spent monthly hours at work)
- time_spend_company (the years spent with the company)
- Work_accident (whether there was a work accident from this employee)
- promotion_last_5years (whether they have had a recent promotion)
- department
- salary (salary level: low/medium/high)
- left (whether (1) or not (0) the employee has left the company)

We decided to take "left" as a response variable $y = \{0,1\}$, such that, predict from the given data whether or no the employee left the company. Thus we're left with 9 explanatory variables, 5 of which are continuous and 4 are categorical. We assumed linear relationship for all explanatory variables, since there was no strong evidence to think otherwise.

II. Initial hypothesis.

Based on some background research (examples: Jensen, 2012; Brown, Thomas, Bosselman, 2015, see Used works section), we hypothesized that people who left would, probably, have:

- lower satisfaction level, which can be a direct reason for leave;
- lower evaluation rating, because employer could lay off an unproductive employee or person may leave by feeling unrecognized;
- higher number of working hours, which can be highly correlated with satisfaction level and be a main reason to leave the company;

- more years (5-7 years) with the company, since new employees are less motivated to find something else;
- higher number of projects, because of longer time with the company;
- have an accident in their record;
- have not been recently promoted;
- have lower salary;
- work at high level stress departments, such as support, sales, accounting, product management departments or at low paid department.

III. Main findings.

The initial hypothesis mainly was confirmed by the following analysis except the assumption about the accidents, evaluation rating, and the number of projects. In general, to distinguish employees who left the company, one needs to consider low satisfaction level, higher evaluation rating, higher number of working hours, approximately 5-7 years with the company for the employees with high satisfaction level and 2-3 years for the employees with low (<0.5) satisfaction level, no accidents, no promotions, low salary. Possible lurking variable is a stress rate, which could be associated with the certain departments, but it was proven only partially for Sales and Support departments.

IV. Data transformation.

Since data was already cleaned, the only one transformation is converting the monthly working hours into the daily working hours for more convenience. For more detailed analysis applied a quadratic term onto # of years with the company.

V. Analyzing the boxplots from the entire dataset (plots are not provided in description, but available in R).

The most significant explanatory variables from the plots appear to be satisfaction level, # of years with the company, since they have less overlapping. Evaluation and # of hours a day are overlapping, but have different means.

Already from the boxplots some parts of our hypothesis were not confirmed. Such as evaluation level for people who left has higher mean and wider spread. Number of projects seems do not influence much the response variable.

VI. Analyzing data from training sample.

A training dataset of 10000 cases was randomly subsetted from the entire dataset. The holdout set was obtained from the rest 4999 cases. The number of y's equal to 1 (left the company) in the training set is 2406 for a proportion 0.2406.

The correlations in the training set of the explanatory variable with the response variables are (Table 1):

x	satisfaction level	evaluation	projects	time in company	daily HRS
r_{xy}	-0.384	0.0025	0.023	0.138	0.071

Table 1. Correlations of explanatory variables in training set

Due to the nature of the response variables (binary) the correlations are relatively small. Nevertheless, the highest absolute correlation values are for satisfaction level and for the time in the company. Thus, we can conclude that those variables are more important.

Negative correlation for the satisfaction level confirms our hypothesis that people with the lower satisfaction level will more likely tend to leave.

For the categorical variables the tables (2 to 5) are the following (**proportions** of people **who left** for each column is the last row):

	Department (employee, count)									
y = left	Acc.	HR	IT	Manag.	Market.	Product mng	Rand	Sales	support	tech
0	369	340	644	382	429	486	432	2085	1080	1347
1	136	147	180	59	131	131	79	679	372	492
prop	0.27	0.3	0.22	0.13	0.23	0.21	0.15	0.25	0.26	0.27

Table 2. Number of people stayed/left by department

	Salary (employee, count)		
y = left	Low	Medium	High
0	3456	3375	763
1	1459	891	56
proportion	0.297	0.209	0.07

Table 3. Number of people stayed/left by salary level

	Accidents (employee, count)	
y = left	0 = no	1 = yes
0	6274	1320
1	2287	119
proportion	0.267	0.08

Table 4. Number of people stayed/left by accidents

	promotions	
y = left	0 = no	1 = yes
0	7391	203
1	2393	13
proportion	0.245	0.06

Table 5. Number of people stayed/left by promotions

The proportion of employees left is almost the same across departments with the exception of Management employees, who have the lowest proportion of left employees equal to 0.13. This is explained by hypothesizing that people on a higher position tend to stay with the company. Sales, support, and accounting departments have slightly higher leaving rates which agrees to our initial hypothesis. Surprisingly, HR department has the highest proportion of employees who left. That can be explained by lower qualifications needed for such type of jobs, such that people tend to quit after gaining higher skills.

For the salary, people with the higher salary seem like to stay. Similar situation for the promotions. Recently promoted employees don't leave the company soon after promotion.

Further analysis with respect to the departments also shows possible explanation, as Sales and Support have the lowest salaries and Management has the highest salaries (Table 6).

	Salary level (employee, count)			Proportion of low salary	Proportion of medium salary
Department	Low	Medium	High		
Accounting	234	226	45	0.46	0.44
HR	219	239	29	0.45	0.49
IT	415	361	48	0.5	0.44
Management	127	155	159	0.28	0.35
Marketing	267	238	55	0.48	0.43
Product mng	303	264	50	0.49	0.43
Random	229	245	37	0.45	0.48
Sales	1433	1160	171	**0.52**	0.42
Support	749	607	96	**0.52**	0.42
Technical	939	771	129	0.51	0.42

Table 6. Salary by department with proportions

Surprisingly, the proportion of people who left after having an accident at workplace isn't high, which contradicts to our initial hypothesis.

VII. Analysis of the boxplots for the training sample.

Boxplots (Figure 1) for the training sample (see below) have the same characteristics (appearance) as the boxplots for the entire dataset (not included into this description), which suggested well randomized pick (training : test = 2:1). The minimum overlap is for the satisfaction level and the number of years with the company. Evaluation and Hours a day have overlapping boxplots, but the mean values are different for each category, which implies certain level of importance of these explanatory variables.

Besides standard boxplots, we decided to plot the most significant explanatory variables – satisfaction level and # of years in company – against each other. The plot shows that people who left the company can be distinguished into three categories:

1) dissatisfied, but working for 4-6 years;
2) somewhat dissatisfied and working for a couple years;
3) satisfied and working 5-7 years;

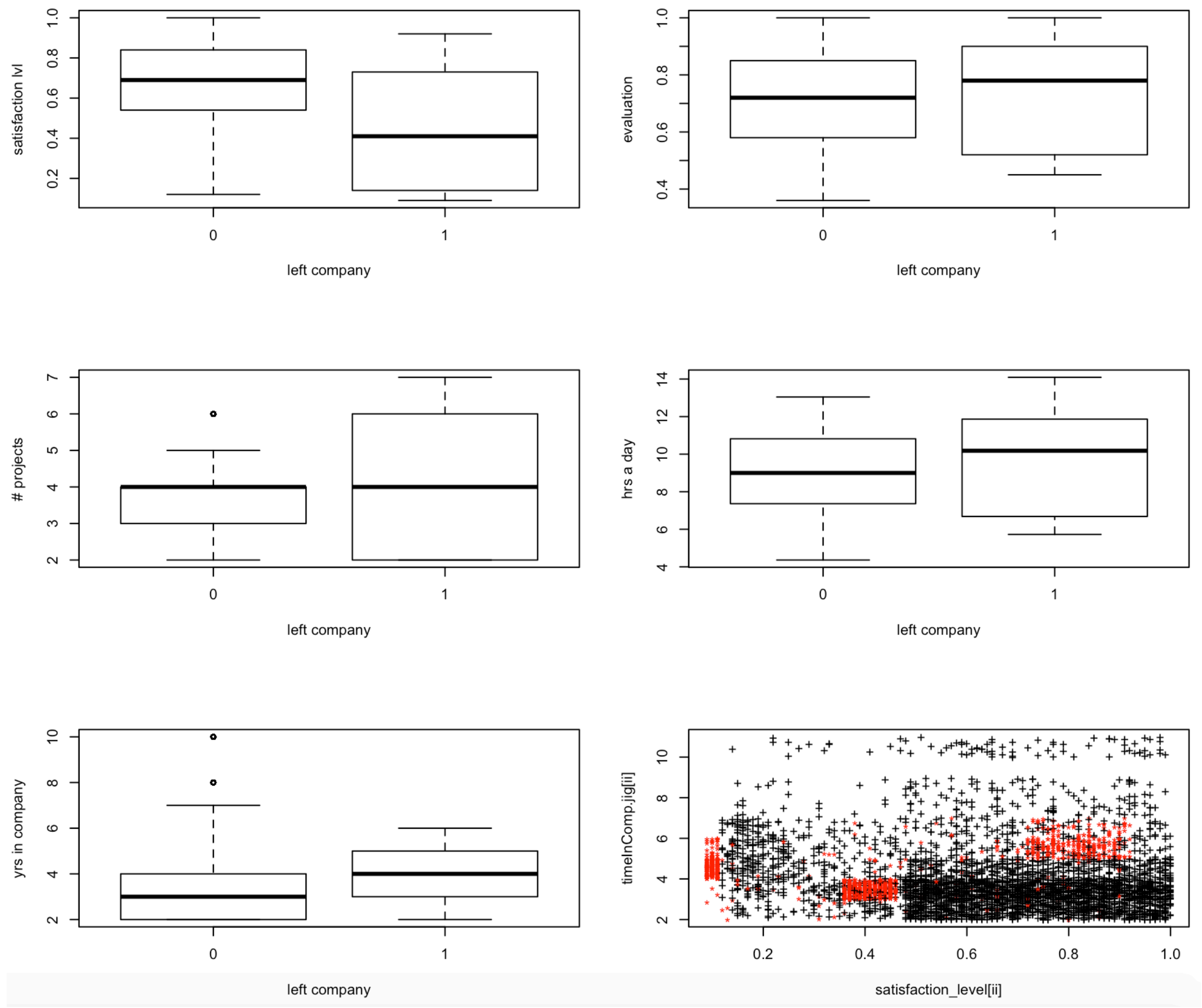


Figure 1. Visual representation of the data for the training subset

The first category was unexpected for our group. We couldn't find a relevant explanation to it. Perhaps, it could be related to a situation, when employee hasn't acquired yet skills/knowledge/experience to move into more desirable position, but still fulfils a "minimum commitment" (Jensen, 2012) to the company.

Second group, most likely, are the employees who didn't like the working environment, so they stood with the company for an average amount of time.

Third group, most likely, people who just desired some changes.

Further analysis (Appendix 1, Figure 6) showed that the low satisfaction level and relatively long period of staying with the company doesn't depend on the type of job. Employees across departments share similar characteristics. People with satisfaction level in a range (0.25, 0.5] tend to leave after 3 years.

Analysis with respect to salary (Appendix 1, Figure 7) also showed that people who left and had lower satisfaction level tended to have low/medium salary. The only department which had more people left with high level of salary and low satisfaction rate is Sales department. That could be explained by more stressful nature of work and bonus dependent wages.

For people who didn’t leave the company, the majority of data appears in the region of short time with the company (2-5 years) and moderate to high level of satisfaction. Which corresponds to our hypothesis, that newer employees know less about the company, have higher satisfaction levels, and less desire to leave.

The analysis across departments with respect to the salary level showed that people with high level of salary tend to stay and have a higher satisfaction level.

VIII. Description of summary statistics of glm.

To fit our model to binary regression, the R function glm is used with option family = binomial. For the dataset including all the variables, the output from R gives the result below (Figure 2).

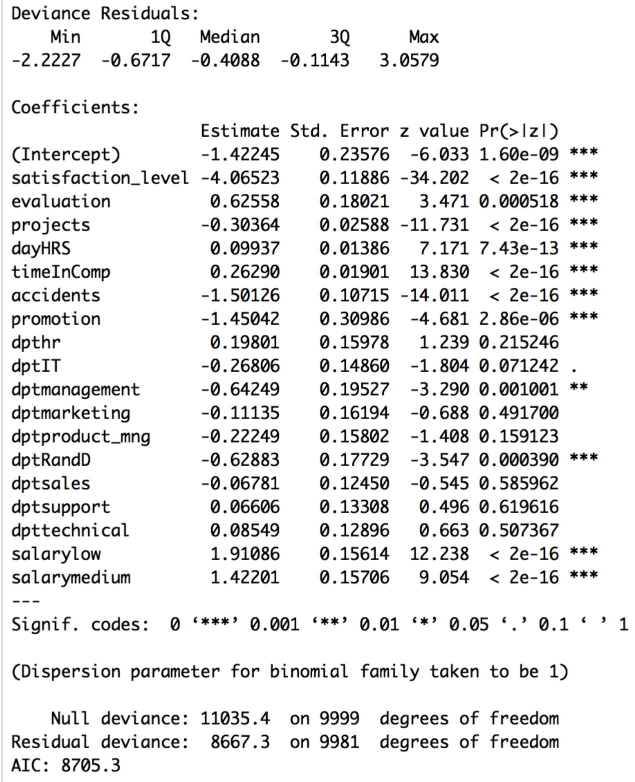
From the output above, we obtained the table of regression coefficients betas, SEs, z-values and P-values, which is similar to multiple regression.

R reports two form of deviance: Null deviance = 11035.4 on 9999 degrees of freedom. Including the independent variables decreased the deviance to 8667.3 on 9981 degrees of freedom.

The null deviance shows how well the response is predicted by the model with nothing but an intercept. The difference between the null deviance and the model’s deviance is distributed as a chi-squared with degrees of freedom equal to the null df minus the model’s df. For our model that would be:

```
> 1-pchisq (11035.4-8667.3, df = (9999-9981))
[1] 0
```

So our Logistic regression model provides an adequate fit for the data.



```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2227  -0.6717  -0.4088  -0.1143   3.0579

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.42245    0.23576  -6.033 1.60e-09 ***
satisfaction_level -4.06523    0.11886 -34.202  < 2e-16 ***
evaluation          0.62558    0.18021   3.471 0.000518 ***
projects           -0.30364    0.02588 -11.731  < 2e-16 ***
dayHRS              0.09937    0.01386   7.171 7.43e-13 ***
timeInComp          0.26290    0.01901  13.830  < 2e-16 ***
accidents          -1.50126    0.10715 -14.011  < 2e-16 ***
promotion          -1.45042    0.30986  -4.681 2.86e-06 ***
dpthr               0.19801    0.15978   1.239 0.215246
dptIT              -0.26806    0.14860  -1.804 0.071242 .
dptmanagement      -0.64249    0.19527  -3.290 0.001001 **
dptmarketing       -0.11135    0.16194  -0.688 0.491700
dptproduct_mng     -0.22249    0.15802  -1.408 0.159123
dptRandD           -0.62883    0.17729  -3.547 0.000390 ***
dptsales           -0.06781    0.12450  -0.545 0.585962
dptsupport          0.06606    0.13308   0.496 0.619616
dpttechnical        0.08549    0.12896   0.663 0.507367
salarylow           1.91086    0.15614  12.238  < 2e-16 ***
salarymedium        1.42201    0.15706   9.054  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 11035.4  on 9999  degrees of freedom
Residual deviance:  8667.3  on 9981  degrees of freedom
AIC: 8705.3
```

Figure 2. GLM with all explanatory variables included.

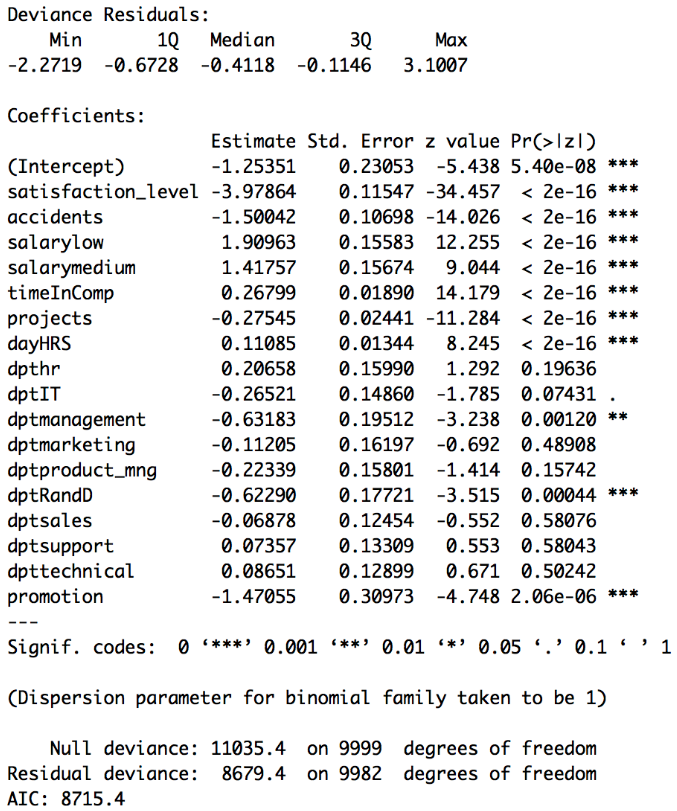
Another test of interest concerns the Null model. We obtained the Null model. Null model assumes the exact “opposite”, in that is assumes one parameter for all of the data points, which means we only estimate 1 parameter. The output of R gives the result below:

```
# Coefficients:
# (Intercept)
# -1.149
#
# Degrees of Freedom: 9999 Total (i.e. Null); 9999 Residual
# Null Deviance:      11040
# Residual Deviance: 11040       AIC: 11040
```

Then we obtained some logistic regression summaries with variable selection based on backward elimination (check R code for the procedure).
The models being compared have:

(a) 9 explanatory variables with k= 19 columns in X because department has 10 categories and salary has 3 categories. Residual deviance: 8667.3, AIC: 8705.3

(b) 8 explanatory variables with k = 18 columns in X after removing evaluation with backward eliminations. Residual deviance: 8679.4, AIC: 8715.4. As shown below (Figure 3):



```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2719  -0.6728  -0.4118  -0.1146   3.1007

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.25351    0.23053  -5.438 5.40e-08 ***
satisfaction_level -3.97864    0.11547 -34.457  < 2e-16 ***
accidents          -1.50042    0.10698 -14.026  < 2e-16 ***
salarylow           1.90963    0.15583  12.255  < 2e-16 ***
salarymedium        1.41757    0.15674   9.044  < 2e-16 ***
timeInComp          0.26799    0.01890  14.179  < 2e-16 ***
projects           -0.27545    0.02441 -11.284  < 2e-16 ***
dayHRS              0.11085    0.01344   8.245  < 2e-16 ***
dpthr               0.20658    0.15990   1.292  0.19636
dptIT              -0.26521    0.14860  -1.785  0.07431 .
dptmanagement      -0.63183    0.19512  -3.238  0.00120 **
dptmarketing       -0.11205    0.16197  -0.692  0.48908
dptproduct_mng     -0.22339    0.15801  -1.414  0.15742
dptRandD           -0.62290    0.17721  -3.515  0.00044 ***
dptsales           -0.06878    0.12454  -0.552  0.58076
dptsupport          0.07357    0.13309   0.553  0.58043
dpttechnical        0.08651    0.12899   0.671  0.50242
promotion          -1.47055    0.30973  -4.748 2.06e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 11035.4  on 9999  degrees of freedom
Residual deviance:  8679.4  on 9982  degrees of freedom
AIC: 8715.4
```

Figure 3. GLM with all except one explanatory variables included.

Model (a), which is still the model with all parameters included, is the best model based on smaller AIC. Model (b) is our second best model. Further improvements to the best model might consist of adding quadratic terms, which is to be done later. Also one of the possible improvement will be merging some department categories to obtain better AIC.

Based on the signs and values of the betas of our best model:

(1) For the categorical explanatory variables, the department category “hr” led to a slightly higher leaving rate than the other categories of department, which is corresponding to our correlation analysis, and the salary category “low” led to a higher leaving rate than the other categories of department. There is a higher leaving rate as salary level decreases even if not all of the betas for the other variables are significant.

(2) For the numerical explanatory variables, there is a higher leaving rate as evaluation increases and as timeInComp increases.

The betas for satisfaction_level, accidents, projects and promotion are all negative, as expected from our initial hypothesis and correlation analysis except for projects and accidents. With higher number of projects, which contradicts with our initial hypothesis, employees are more likely to stay in the company, one possible reason is that employees feel motivated with sense of accomplishment after completing more projects. The reasoning about the accidents is explained in the correlation section.

The betas for evaluation, timeInComp and dayHRS are positive. The betas for timeInComp and dayHRS are corresponding to our initial hypothesis and correlation analysis, however, with higher evaluation level, employees tend to leave the company. One guess of the possible reason is that with higher evaluation level, employees are more likely to find a more valuable position in a better company.

(3) For the interpretation of betas, the slope for timeInComp in the logistic regression is 0.26290. Since $e^{0.26290}$ = 1.301, and $e^{0.26799*2}$ = 1.692, the increased odds of leaving the company for someone in this population is 1.301 for an increased year of 1 and 1.692 for an increased year of 2.

The slope for evaluation level in the logistic regression is 0.62558. Since $e^{0.62558*1}$ = 1.065, and $e^{0.62558*0.5}$ = 1.367, the increased odds of leaving the company for someone in this population is 1.065 for an increased evaluation level of 0.1 and 1.367 for an increased evaluation level of 0.5.

The slope for dayHRS in the logistic regression is 0.09937, Since $e^{0.09937}$= 1.104 and $e^{0.11085*5}$ = 1.644, the increased odds of leaving the company for someone in this population is 1.104 for an increased daily working hours of 1 and 1.644 for an increased daily working hours of 5.

In addition to the full model with linear terms, we’ve decided to check for improvements by applying quadratic terms to each of the continuous explanatory variables (see R file for more details).

The best model (based on AIC) appeared to be after applying a quadratic term towards years in company. However, after using this model, we’ve got a perfect separation for the model. That suggests possible usage of hidden logistic regression model, which was out of scope of our project. The SE of betas didn’t appear to be large, which means there is not a strong multicollinearity between the explanatory variables. Perhaps, quadratic term just overloaded the rest of the data. Thus for the further analysis, we’ve decided to proceed with the best regular terms model to avoid inconsistency.

We also tried to combine several departments in a meaningful way, then to apply the model with quadratic term, but it almost didn’t affect AIC (Appendix 1, Figure 9) . Possible explanation to that, that the data is consistent across all departments, such that department explanatory variable becomes irrelevant for the model.

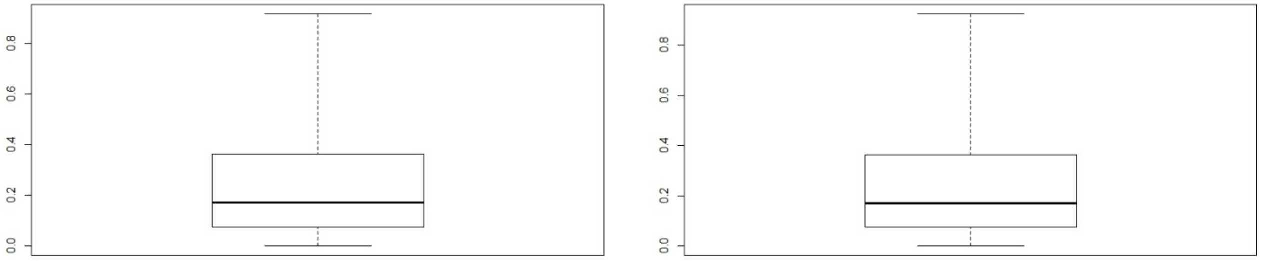
IX. Misclassification.

After we analyzed the glm summary statistics, we used in-sample and out-sample misclassification evaluation to compare the fit of the two models.

For in-sample misclassification, Summary statistics of $\widehat{\pi_i}$, the predicted probability for an employee to leave the company, are in the following Table 7 (please refer to R code to see the sequence of actions to find these values).

Model	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Fit1	0.000999	0.0752	0.171	0.241	0.362	0.915
Fit2	0.000923	0.0759	0.170	0.241	0.363	0.924

Table 7. Summary statistics for in-sample misclassification



Fit1 boxplot

Fit2 boxplot

Figure 4. Boxplots for in-sample misclassification for Fit1 and Fit2 models.

From the output of the summary statistics we can see that:

1. Mean of $\widehat{\pi_i}$ for both fit1 and fit2 is 0.241, which is same as the proportion of y's equal 1 in the training set of 10000 cases
2. Quantile data and box plots show their distribution are quite similar too.

Summaries are given below for the in-sample and out-of-sample misclassification when thresholds (τ) are 0.5, 0.3, and 0.1. Based on the definition of misclassification, the smaller the threshold, the higher possibility the prediction will be true, and the higher possibility of false positive. Note, that in the tables a model with the larger misclassification rate is marked red.

Table 8. In-sample table of misclassification:

	$\hat{\pi} \leq 0.5$	$\hat{\pi} \geq 0.5$	misclass	$\hat{\pi} \leq 0.3$	$\hat{\pi} \geq 0.3$	misclass	$\hat{\pi} \leq 0.1$	$\hat{\pi} \geq 0.1$	misclass
$y = 0$ (fit1)	7042	552	0.0727	6108	1486	0.196	3154	4440	0.585
$y = 1$ (fit1)	1575	831	0.655	754	1652	0.313	91	2315	0.0378
$y = 0$ (fit2)	7039	555	0.0731	6124	1470	0.194	3140	4454	0.587
$y = 1$ (fit2)	1545	861	0.642	755	1651	0.314	96	2310	0.0399

Table 9. Out-of-sample table of misclassification:

	$\hat{\pi} \leq 0.5$	$\hat{\pi} \geq 0.5$	misclass	$\hat{\pi} \leq 0.3$	$\hat{\pi} \geq 0.3$	misclass	$\hat{\pi} \leq 0.1$	$\hat{\pi} \geq 0.1$	misclass
$y = 0$ (fit1)	3578	256	0.0668	3131	703	0.183	1645	2189	0.571
$y = 1$ (fit1)	737	428	0.633	364	801	0.312	46	1119	0.0394
$y = 0$ (fit2)	3571	263	0.0686	3130	704	0.184	1639	2195	0.573
$y = 1$ (fit2)	725	440	0.622	363	802	0.311	48	1117	0.0412

Based on the tables of misclassification, we observed that as the threshold increases the number of false positive cases decreases and the number of false negative cases increases. This observation is expected as it illustrates the trade-off in the two misclassification rates as the threshold changes.

Overall, fit1 is marginally better than fit2 for both in-sample and out-of-sample misclassification rates when the threshold is 0.1. When the threshold is 0.5 and 0.3, there is no strong evidence that shows one model is significantly better than the other.

X. Calibration check.

The final step of our logistic regression modeling is to perform Hosmer-Lemeshow calibration check to see how well our model fits the data. We apply corresponding procedures to the two fitted models we have been analyzing:

1st model: All parameters included: fit1 = full.glm

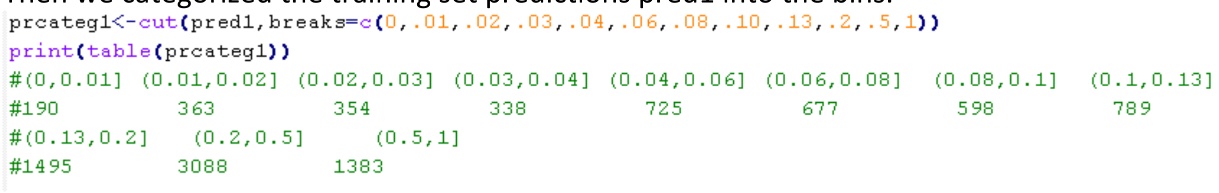
2nd model: evaluation variable taken out. fit2 <- glm (left ~ satisfaction_level + accidents + salary + timeInComp + projects + dayHRS + dpt + promotion, data=train, family="binomial")

To evaluate fit1:

We randomly picked 11 bins for predicted probabilities (of quitting).

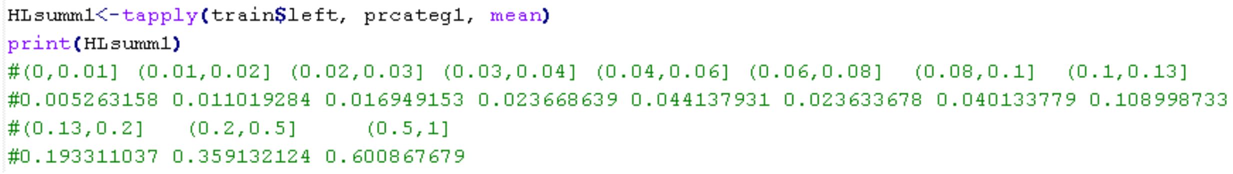
```
c(0,.01,.02,.03,.04,.06,.08,.10,.13,.2,.5,1)
```

Then we categorized the training set predictions pred1 into the bins.



```
prcateg1<-cut(pred1,breaks=c(0,.01,.02,.03,.04,.06,.08,.10,.13,.2,.5,1))
print(table(prcateg1))
#(0,0.01] (0.01,0.02] (0.02,0.03] (0.03,0.04] (0.04,0.06] (0.06,0.08]  (0.08,0.1]  (0.1,0.13]
#190         363         354         338         725         677         598         789
#(0.13,0.2]   (0.2,0.5]     (0.5,1]
#1495        3088        1383
```

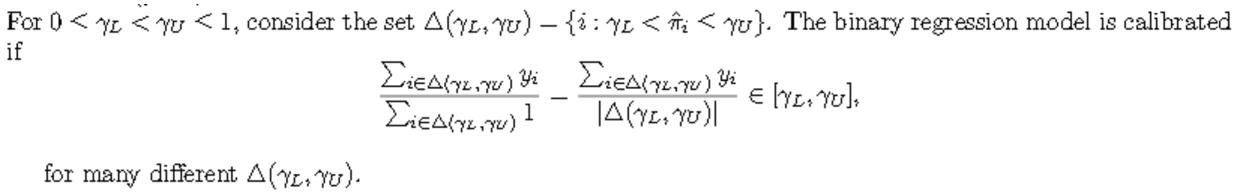
After that we put the corresponding observed values into the bins, and calculate mean (expected value) for each of the bin.



```
HLsumm1<-tapply(train$left, prcateg1, mean)
print(HLsumm1)
#(0,0.01] (0.01,0.02] (0.02,0.03] (0.03,0.04] (0.04,0.06] (0.06,0.08]  (0.08,0.1]  (0.1,0.13]
#0.005263158 0.011019284 0.016949153 0.023668639 0.044137931 0.023633678 0.040133779 0.108998733
#(0.13,0.2]   (0.2,0.5]     (0.5,1]
#0.193311037 0.359132124 0.600867679
```

Intuitively, if the prediction fit the data well, for each bin, the expected values of the true probabilities would be same as that of the predicted probabilities. So we want as many mean values calculated fall within the lower and upper boundaries of their corresponding bins as possible.

That is we want:



For $0 < \gamma_L < \gamma_U < 1$, consider the set $\Delta(\gamma_L, \gamma_U) = \{i : \gamma_L < \hat{\pi}_i < \gamma_U\}$. The binary regression model is calibrated if

$$\frac{\sum_{i\in\Delta(\gamma_L,\gamma_U)} y_i}{\sum_{i\in\Delta(\gamma_L,\gamma_U)} 1} = \frac{\sum_{i\in\Delta(\gamma_L,\gamma_U)} y_i}{|\Delta(\gamma_L,\gamma_U)|} \in [\gamma_L, \gamma_U],$$

for many different $\Delta(\gamma_L, \gamma_U)$.

For this fit1 above, we see that among the 10 bins, 7 bins contain the true expected value.

Perform the same procedure for fit2. And we found 6 bins contain the true expected value.

Therefore, we could see that fit1 is better calibrated compared with fit2. This result aligns with the other test we did and mentioned previously.

Notice the Hosmer-Lemeshow test sometimes is not considered most accurate due multiple reasons, including the fact that we are picking the number of bins and bin size randomly. Therefore, it is usually recommended to do multiple repetitions of the same test with different bin divisions. So, we did the same test for the following divisions and got the following results:

Bin division	Num-containing bin in fit1 vs fit2
c(0,.02,.04,.06,.08,.12,.2,.4,.6,1)	Fit1 same as Fit2 (5 vs 5 inclusive bins)
c(0,.02,.04,.07,.09,.15,.2,.4,.6,1)	Fit1 better than Fit2 (6 vs 4 inclusive bins)

Table 10. Results of multiple repetition of the test

We also considered the possibility that introducing quadratic terms may improve the calibration. However intuitively, the explanatory variables we had did not seem to be possible to have quadratic relationship with quitting probability. That said, we still tried fitting with quadratic term and no significant improvement was observed. We conclude maybe the non-monotonicity is within sampling variability.

XI. Limitations and further research.

Our model is relatively simple and provides only a binary response for the question. However, for more practical reasons, more sophisticated model would be more preferable. For instance, ability to predict how long the person will stay with the company. But this was out of scope of our research.

Since our best model included all variables, we found it redundant to look for a better AIC. However, improvement in AIC score and, therefore improvement in model could be achieved by merging several departments or running cross-validation tests in smaller batches.

Additional limitation is lack of good explanatory variables. As it was mentioned in findings, some variable as stress level would improve the model.

The last one is using different tools, like PCA or kernel PCA, to define a better model and find better explanation for the relations between explanatory variables and the response variable.

XII. Used works:

1. Jensen, G. D. “Why Good People Leave Good Jobs.” Science AAAS. 17 Aug 2012. DOI: 10.1126/science.caredit.a1200093
2. Brown, E.A., Thomas, N.J., Bosselman, R.H. (2015) Are they leaving or staying: A qualitative analysis of turnover issues for Generation Y hospitality employees with a hospitality education. *International Journal of Hospitality Management*, 46, 130-137. http://dx.doi.org.ezproxy.library.ubc.ca/10.1016/j.ijhm.2015.01.011

XIII. Contribution section.

Our team was formed partially via piazza, partially via personal connections. Initially, we didn't have strong preferences for the project idea and decided on it later after analyzing the available data sources, such as Kaggle, Google statistics, etc. The topic was chosen based on the relative interest of all the team members in it. After it was done, we had several discussions via the chat and set up the team folder on Google docs to share the results of individual work. We've worked on the proposal using that on-line folder.

From received feedback, we've decided to change type of response variables and stopped on the binary response model. Through the work on the final part of the project we had several on-line discussions and in person meetings to agree on the steps.

Project contributors are listed in alphabetical order by the last name:

Aleksandra Budkina: proposing the topics, drafting and finalizing proposal, writing the majority of R code for the final product; writing project description, Section 1, Section 2, Outline, Contribution section for the final project submission.

Jason Chau: main contributor in topic choice, reviewing and improving proposal, didn't participate in final submission.

Woo Lee: reviewing and improving proposal, writing description on misclassification rate for the final project submission, writing R code for running project in batch, reviewing and improving final submission.

Wenyue Li: reviewing and improving proposal, writing GLM description for the final submission, reviewing and improving final submission.

Ke Miao: proposing the topics, reviewing and improving proposal, writing R code and description on calibration check, writing R code on quadratic terms, writing description on misclassification rate for the final project submission, reviewing and improving final submission.

Appendix 1

Analysis of Satisfaction level vs Time in Company and Satisfaction level vs Salary across departments.

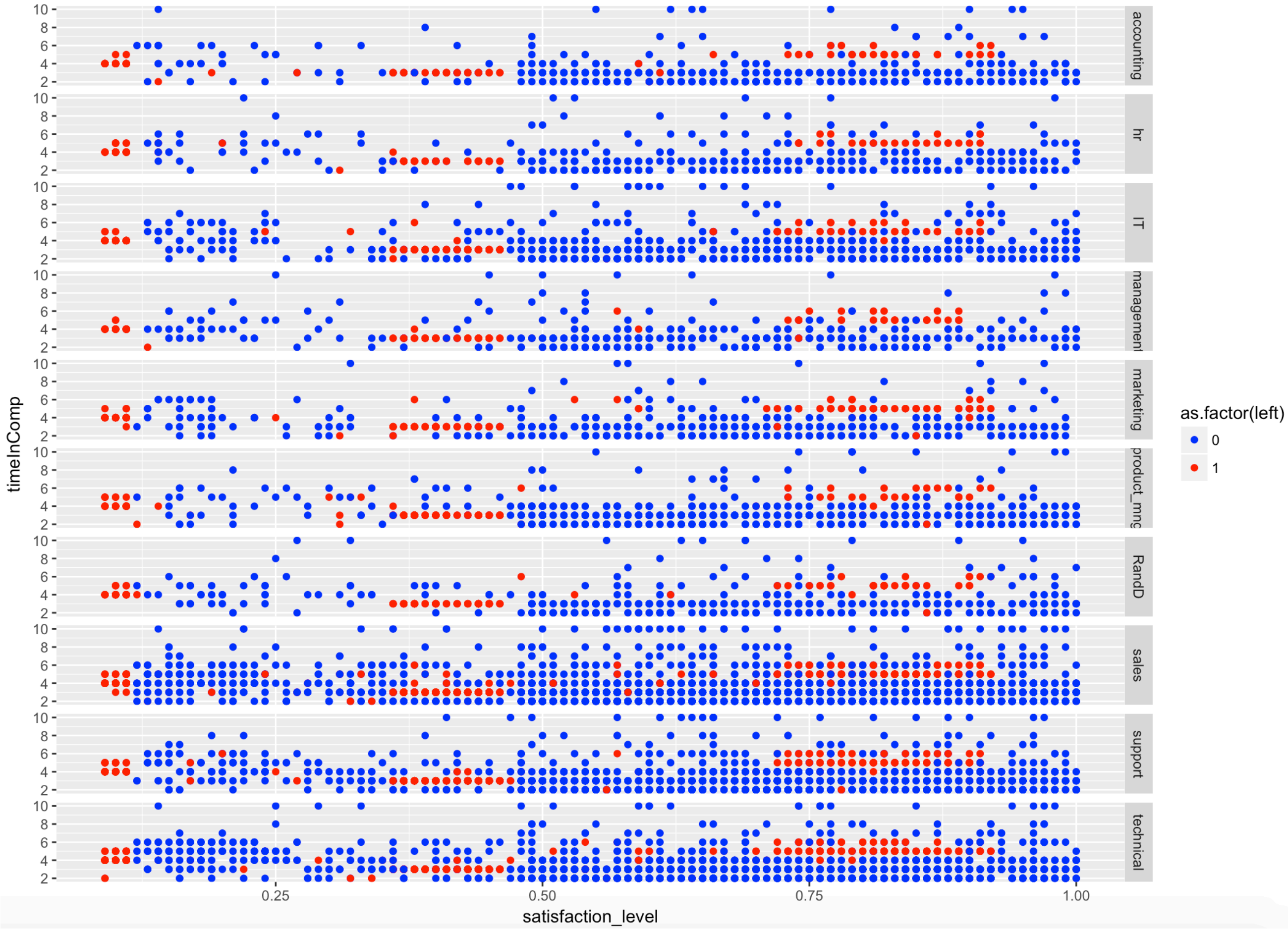


Figure 6. Analysis of Satisfaction level vs Time in Company across departments

We can see that no employees stayed with the satisfaction level < 0.25.

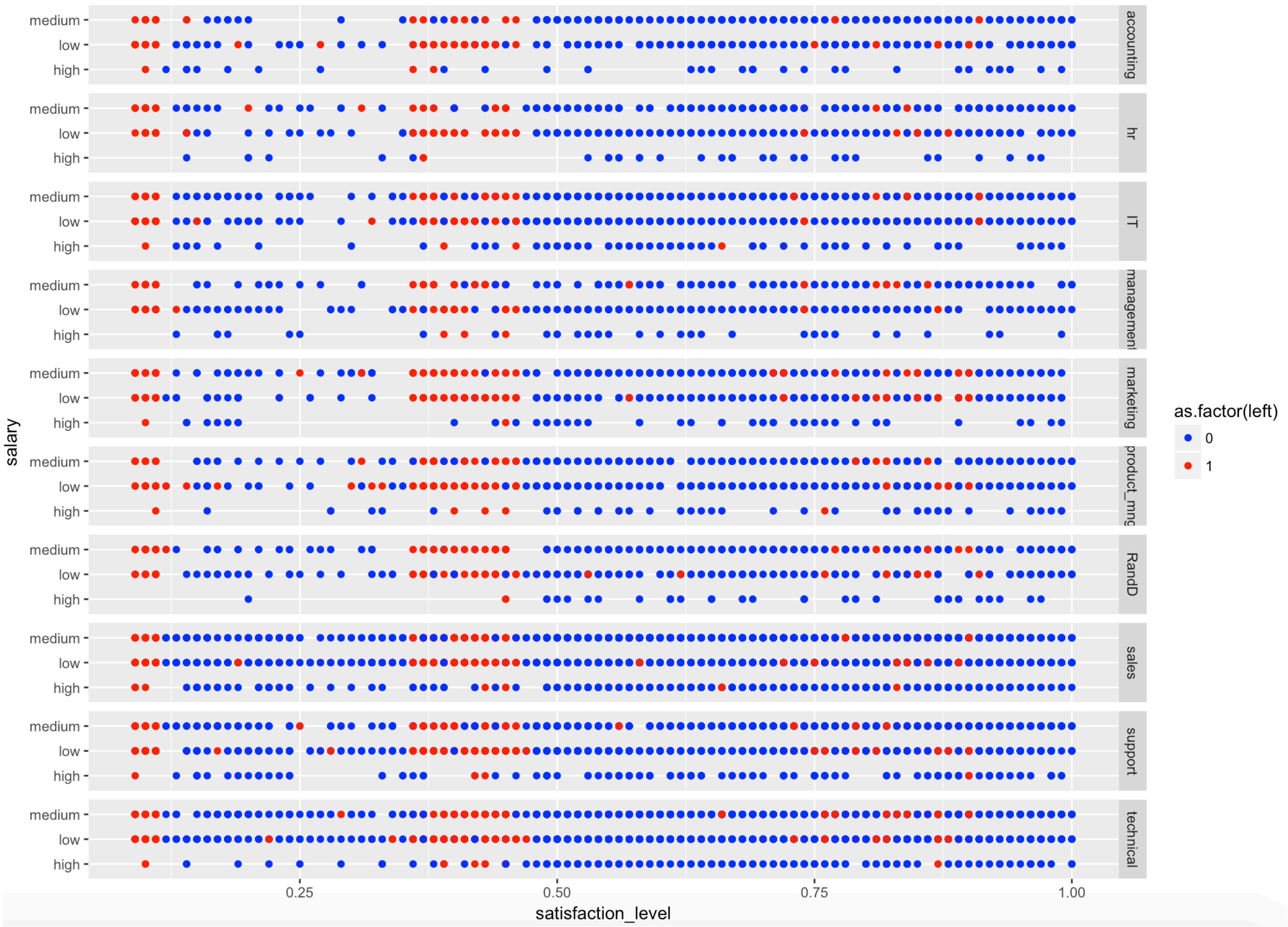
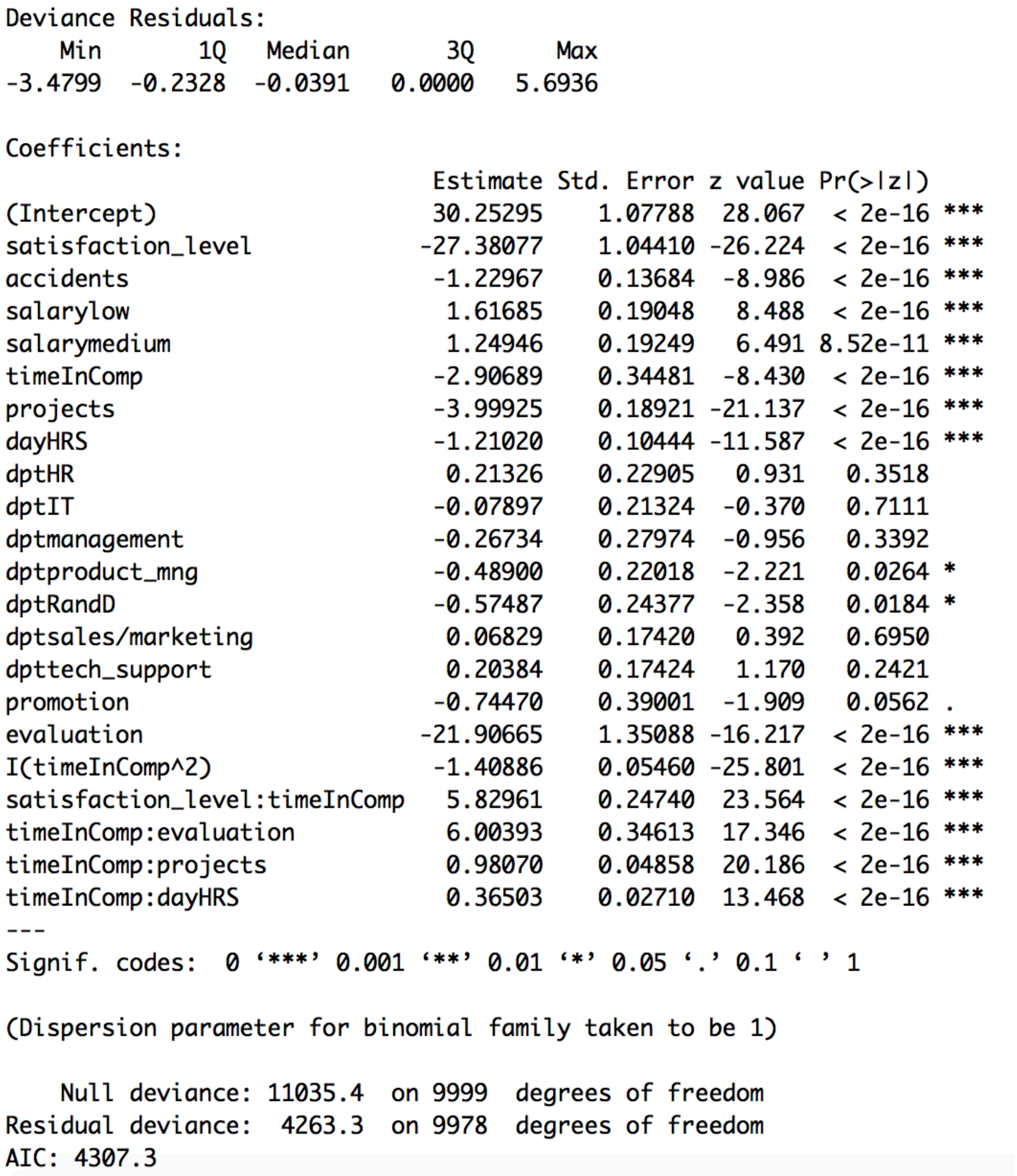


Figure 7. Analysis of Satisfaction level vs Salary across departments

There are more people among who left the company with low/medium range of salary.



```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.4799  -0.2328  -0.0391   0.0000   5.6936

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                   30.25295    1.07788  28.067  < 2e-16 ***
satisfaction_level           -27.38077    1.04410 -26.224  < 2e-16 ***
accidents                     -1.22967    0.13684  -8.986  < 2e-16 ***
salarylow                      1.61685    0.19048   8.488  < 2e-16 ***
salarymedium                   1.24946    0.19249   6.491 8.52e-11 ***
timeInComp                    -2.90689    0.34481  -8.430  < 2e-16 ***
projects                      -3.99925    0.18921 -21.137  < 2e-16 ***
dayHRS                        -1.21020    0.10444 -11.587  < 2e-16 ***
dptHR                          0.21326    0.22905   0.931   0.3518
dptIT                         -0.07897    0.21324  -0.370   0.7111
dptmanagement                 -0.26734    0.27974  -0.956   0.3392
dptproduct_mng                -0.48900    0.22018  -2.221   0.0264 *
dptRandD                      -0.57487    0.24377  -2.358   0.0184 *
dptsales/marketing             0.06829    0.17420   0.392   0.6950
dpttech_support                0.20384    0.17424   1.170   0.2421
promotion                     -0.74470    0.39001  -1.909   0.0562 .
evaluation                   -21.90665    1.35088 -16.217  < 2e-16 ***
I(timeInComp^2)               -1.40886    0.05460 -25.801  < 2e-16 ***
satisfaction_level:timeInComp  5.82961    0.24740  23.564  < 2e-16 ***
timeInComp:evaluation          6.00393    0.34613  17.346  < 2e-16 ***
timeInComp:projects            0.98070    0.04858  20.186  < 2e-16 ***
timeInComp:dayHRS              0.36503    0.02710  13.468  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 11035.4  on 9999  degrees of freedom
Residual deviance:  4263.3  on 9978  degrees of freedom
AIC: 4307.3
```

Figure 8. GLM summary after applying a quadratic term towards years and combining several departments.