HR Analytics - Why Do Employees Leave Prematurely?

By Abishek Ganesh, Prashanthi Chandrasekaran, Sourabh Mahajan and Sushidhar Jayaraman

Abstract

One of the main issues that companies these days are facing is the issue of Employee attrition. This could be due to many factors like the number of hours an employee works, satisfaction level, salary, promotions etc. Even though there are several such factors, job satisfaction was found to be the main reason for employee attrition. Job satisfaction could mean different things to different people. For some, it is whether they like the job and for some it is how content they are with the job. In this paper, our main objective is to predict the satisfaction level of employees based on various parameters like number of projects the employee has worked in, the average number of hours per month that he works, salary etc. For this purpose, we have used linear regression.

Introduction

Great employees are an organizations best resources. To retain such employees who are a great asset to the company, job satisfaction is of at most importance. This is because satisfied employees are much less likely to leave. Since companies spend a lot of time and energy in hiring their employees, it is very important for them to retain them. For this, they could perform some analysis to see which of their employees are at a risk of leaving the company. This would help them to take measures to retain their talented trained employees. The main purpose of this study is to predict the level of job satisfaction. There are several factors that lead to job satisfaction of employees. Salary for example, employees with high salary are not the ones with high job satisfaction. On the other hand, low and medium salary people are the ones who are not

satisfied with their jobs and hence tend to jump to another company. Another major factor is the average number of hours per month that an employee spends in the company. People who tend to spend more than 200 hours on an average, seem to be less satisfied with their jobs than the ones that spend lesser than 200 hours. In this paper, we have used linear regression to predict the satisfaction level of employees.

Literature Review

The Initial observation is essential since any empirical model exploring the effects of promotions on job satisfaction must provide explanation for salary. It likewise recommends a mechanism through which job promotions may impact employee job satisfaction: promotions raise the employee to a higher position with respect to the individuals who don't get one. But in current data set, by running various models, we find that promotion in last five years is not significant in predicting the employee job satisfaction.

As an indirect measure of the connection between employee job satisfaction and employee turnover, several papers have also explored the significance of employee job satisfaction with advancement in opportunities on future job, with blended outcomes. “Clark (2001) finds that both satisfaction with pay and job security are the most important job satisfaction categories for determining future quits, while satisfaction with promotion opportunities is not a significant factor”. Using data on British nurses, “Shields and Ward (2001) find that dissatisfaction with promotion and training opportunities have a stronger effect on intentions to quit than dissatisfaction with workload or pay”. After analyzing the results for our data, we can say there is a negative relation between job satisfaction and employee left the company. Also, this relation depends upon salary of employee and department of the employee in which he works.

Considering the studies that states the relationship between job satisfaction and organizational performance, the number of studies focused on this relation is much smaller in connection to the number of studies dealing with the relationship between job satisfaction and individual performance. These studies are new, and give changing results. Some authors have discovered a positive correlation between job satisfaction and individual performance. On the other hand, some authors do not find any statistically significant correlation between these two variables. In our case, we are focusing on relation between job satisfaction and performance (last evaluation) of a single employee. Also, our performance attribute tells about the period for last evaluation when it has happened. We find that there is a negative relationship between job satisfaction and last evaluation. It means if the time increases for the last evaluation for an employee, there will be decrease in job satisfaction. But last evaluation has diminishing effect on the job satisfaction.

Thus, although clear directions of causality in the association between job satisfaction and last_evaluation stay uncertain, preliminary evidence has suggested that employee attitudes have associations with performance outcomes and the other way around; the performance is correlated to job satisfaction. All the proof suggests that the association between employee's job satisfaction and last_evaluation is complex and insufficiently researched.

Data

The project uses a dataset from Kaggle where a company has been having trouble retaining their best and most experienced employees. The data frame consists of 14999 observations of 10 variables, which are:

Variables	**Description**	**Median**	**Std.Dev**
satisfaction_level	Employees overall job satisfaction level based on a survey. Ranges between 0 and 1.	0.64	0.25
last_evaluation	The grade the employees got at their last evaluation by their manager. Ranges between 0 and 1.	0.72	0.17
number_projects	The number of projects the employee has been involved in.	4.00	1.23
average_monthly_hours	The number of monthly hours the employee is working.	200.00	50.00
time_spent_company	The number of years the employee has been working for the company	3.00	1.46
work_accident	Whether the employee has already had a work	0.14	0.35

	accident in the past (1 for yes, 0 for no)		
promotion_last_5_years	Whether the employee has been promoted (1 for yes, 0 for no)	0.02	0.144
sales	The department the employee is working for	N/A	N/A
salary	The current level of salary of the employee (3 categories: high, medium, low)	N/A	N/A
left	Whether the employee has left. 0 if the employee is still working for the company, 1 if not.	N/A	N/A

The first thing to look at is how bad the turnover of an employee in this company. Assuming each record corresponds to an employee, the company has 14999 employees and the turnover rate is 24%. The turnover rate is usually calculated on an annual basis and we don't what period

this dataset covers exactly. But any company of this size with an R&D department, turnover rate of 24% is high.

The satisfaction level seems to be highly correlated with the people who had left the company. Considering the satisfaction score less than 0.5 as low, more number of employees had left the company who has low satisfaction level *(Fig 1*).

By breaking down into department, HR and accounting have low median satisfaction levels and high turnover rate whereas R&D and Management have higher satisfaction levels and lower turnover rate (*Fig 2*).

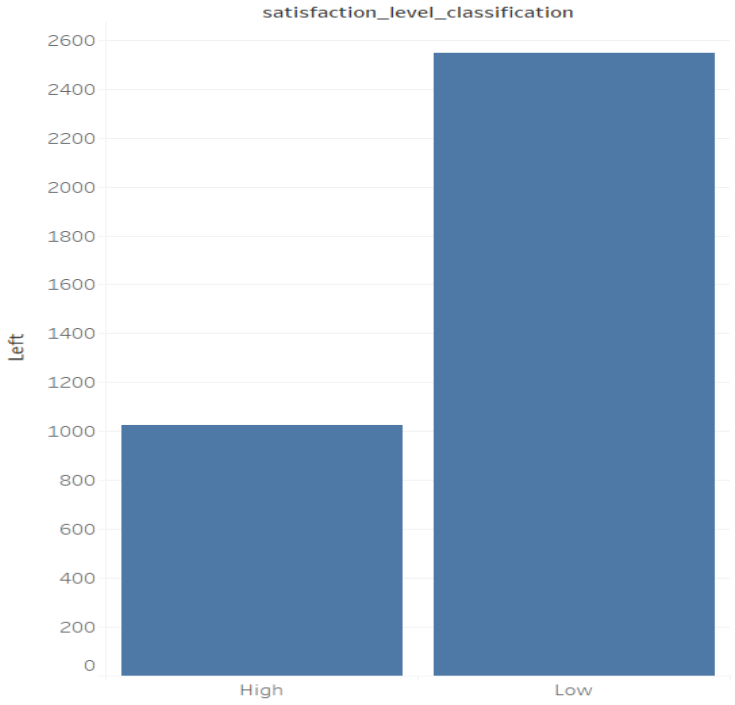

Fig 1: Satisfaction level vs Left

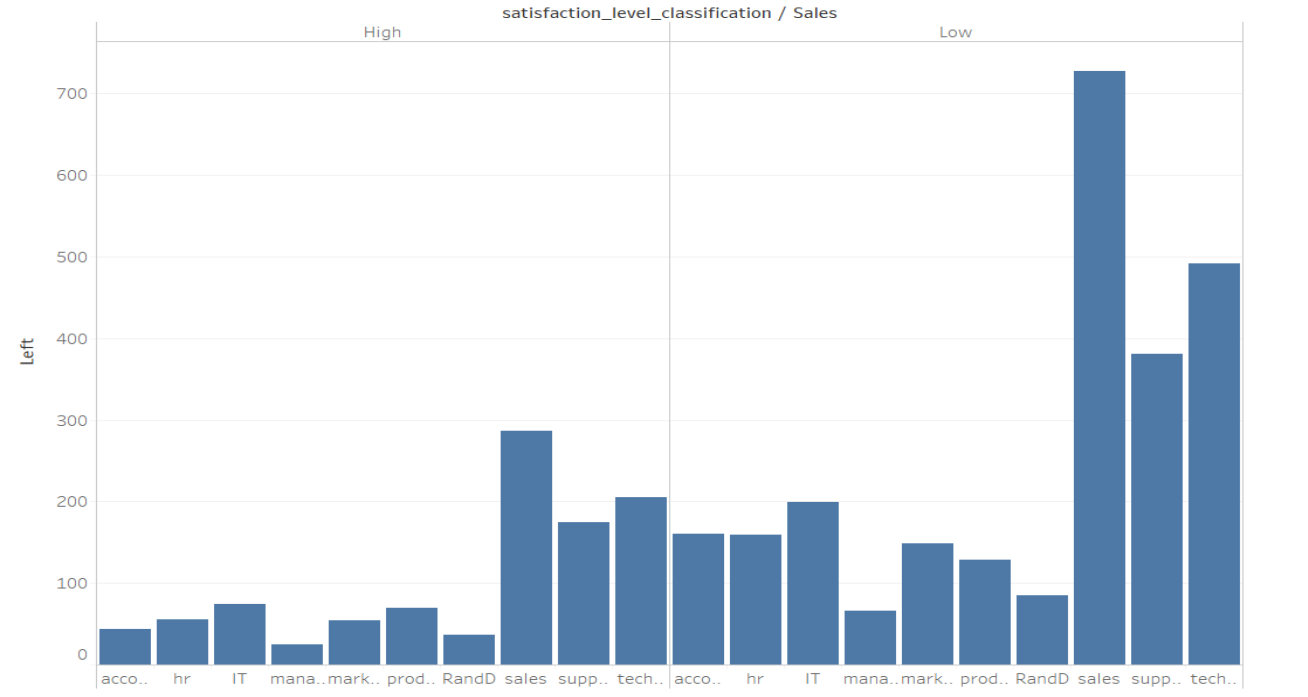

Fig 2: Satisfaction level / Sales vs Left

The main objective is to determine why the best and most experienced employees leave the company. But this dataset does not explicitly mention who are the best employees. We have decided to use performance evaluation score to determine high performers. As the following graph (*Fig 3*) shows, employees with lower performance scores as well as higher performance scores tend to turnover, and people with average performance scores are less likely to leave. The median score is 0.72 and employees with 0.80 are considered as high performers and employee with score less than 0.56 are considered as low performers.

Within high performers, we are interested in most experienced employee. Adding time spent at the company on top of the performance evaluation score, we could see that high performers with 4 to 6 years of experience are at the high of risk of leaving the company (*Fig 4*).

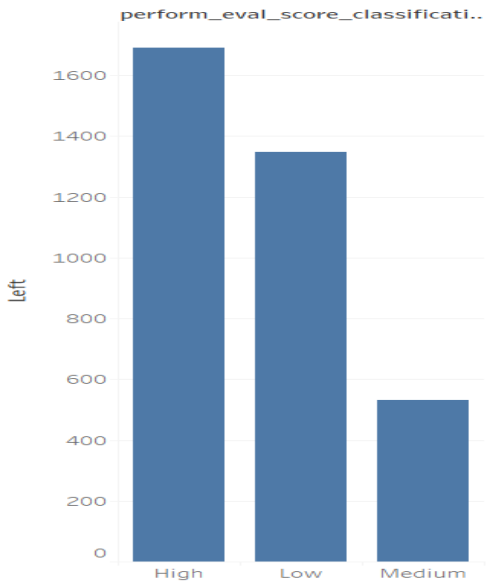

Fig 3: Performance Evaluation vs Left

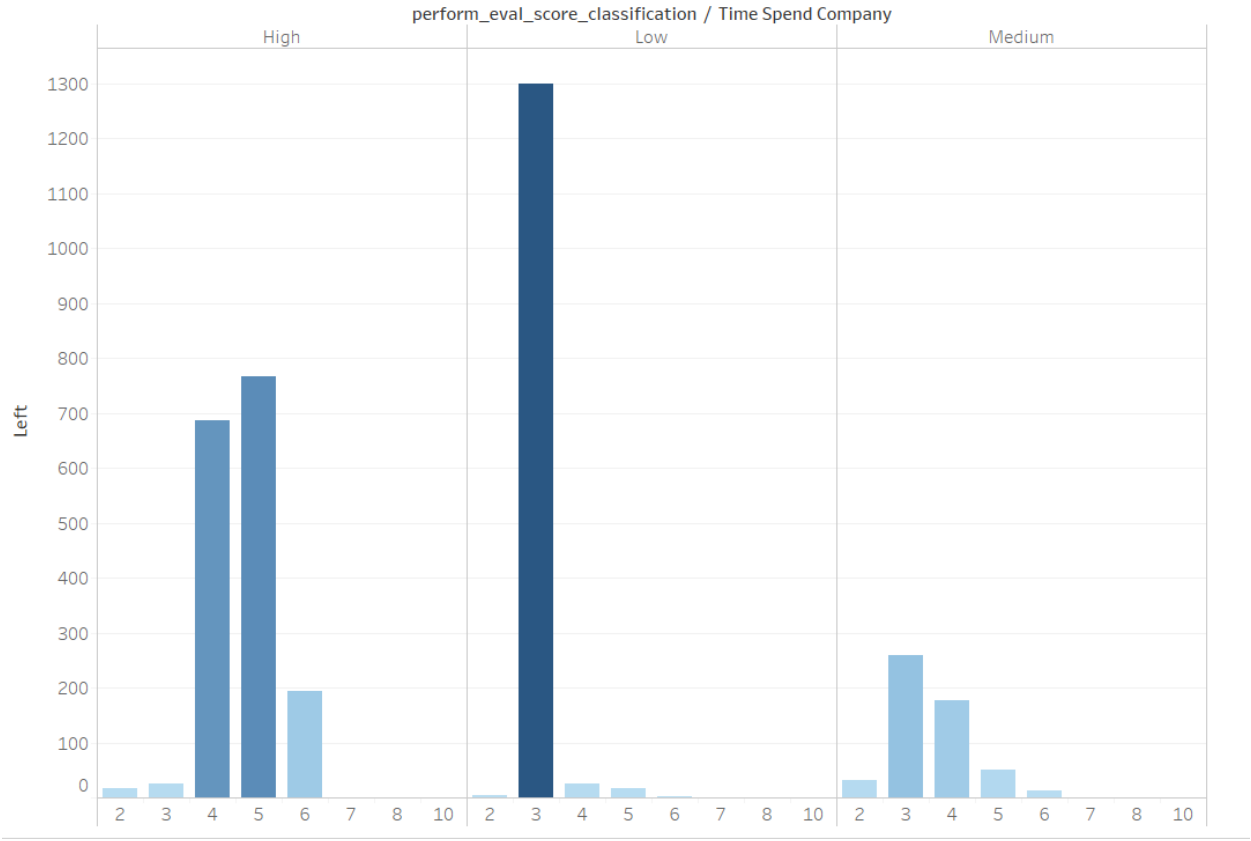

Fig 4: Performance Evaluation / Time Spend Company vs Left

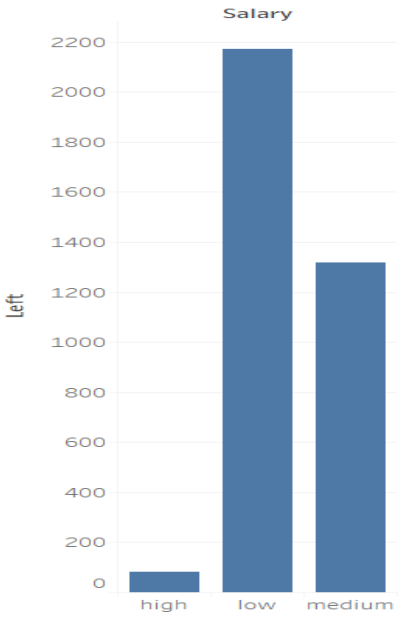

Fig 5: Salary vs Left

There is no big surprise with respect to salary as people with low and medium salary leave the company while people with high salary are staying (Fig 5).

Following correlation chart (*Fig 6*) shows that employees who have higher satisfaction level leave the organization when they work on more number of projects simultaneously or more number of hours each month and when they have already spent a long time in the company. Leave decisions are also influenced by a low salary level and when employees haven't got a promotion during the last 5 years. Also, we could see that evaluation score is positively correlated with spending more monthly hours and number of projects.

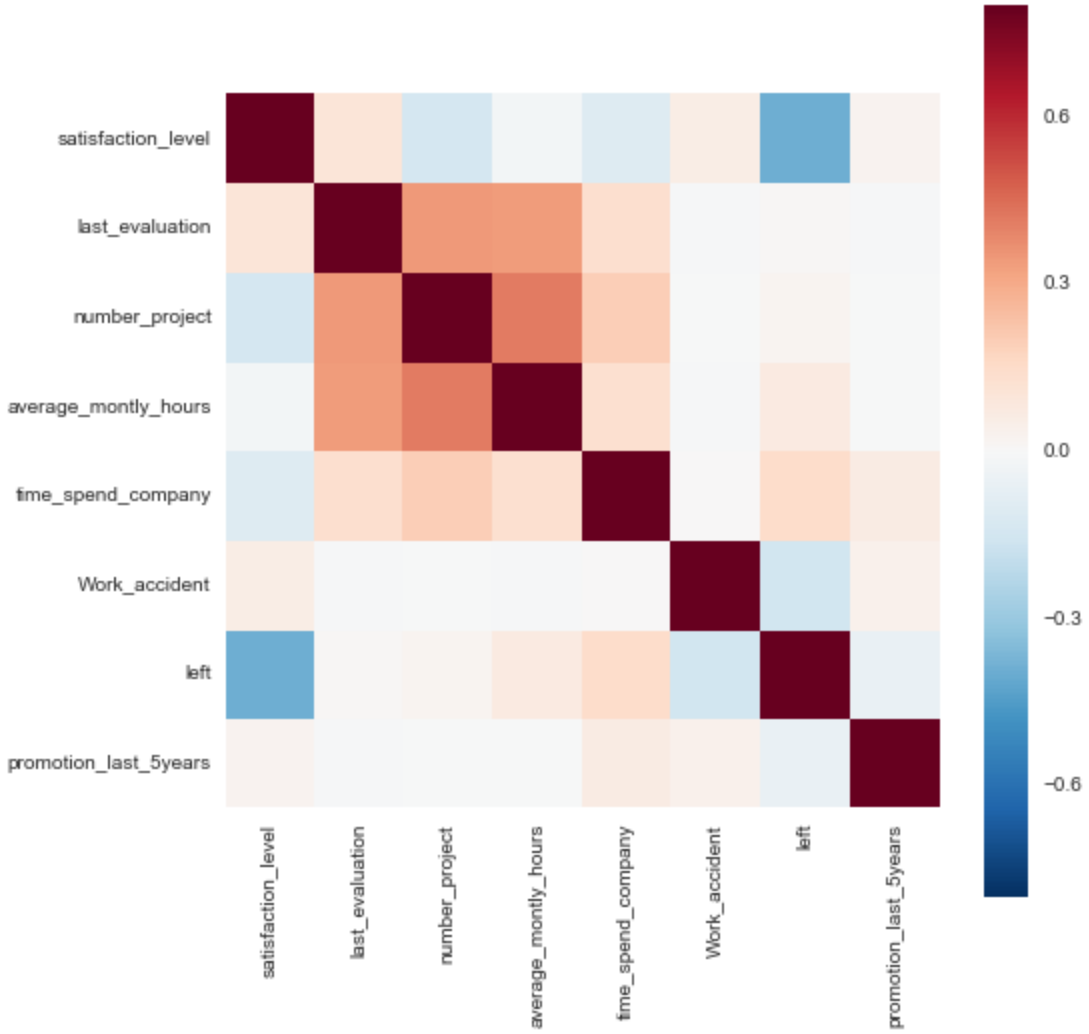

Fig 6: Correlation Chart

Results:

y = data$satisfaction_level

x1 = data$last_evaluation

x2 = data$number_project

x3 = data$average_montly_hours

x4 = data$time_spend_company

x5 = data$Work_accident

x6 = data$left

x7 = data$promotion_last_5years

x8 = data$sales

x8 = dummy(x8)

x9 = data$salary

x9 = dummy(x9)

```
> res = lm(y~x1 + x2 + x3 + x4 + x5 + x6+ x7+x8+x9) # Selected
> summary(res)

Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9)

Residuals:
     Min       1Q   Median       3Q      Max
-0.64765 -0.13755 -0.01158  0.17083  0.52220

Coefficients: (2 not defined because of singularities)
                  Estimate  Std. Error t value             Pr(>|t|)
(Intercept)       0.62800883  0.01087327  57.757 < 0.0000000000000002 ***
x1                0.24593100  0.01167675  21.062 < 0.0000000000000002 ***
x2               -0.04086581  0.00169250 -24.145 < 0.0000000000000002 ***
x3                0.00019078  0.00004126   4.624              0.0000038 ***
x4               -0.00547274  0.00130348  -4.199              0.0000270 ***
x5               -0.00022928  0.00523927  -0.044                0.96510
x6               -0.22407490  0.00445831 -50.260 < 0.0000000000000002 ***
x7                0.01076749  0.01283275   0.839                0.40145
x8x8accounting   -0.02389536  0.00910761  -2.624                0.00871 **
x8x8hr           -0.00733350  0.00924877  -0.793                0.42784
x8x8IT            0.00158880  0.00766167   0.207                0.83572
x8x8management   -0.00589802  0.01016029  -0.580                0.56159
x8x8marketing     0.00102403  0.00874241   0.117                0.90676
x8x8product_mng   0.00312819  0.00855995   0.365                0.71478
x8x8RandD        -0.01015662  0.00903020  -1.125                0.26072
x8x8sales         0.00338131  0.00550530   0.614                0.53910
x8x8support       0.00539307  0.00636403   0.847                0.39677
x8x8technical             NA          NA      NA                    NA
x9x9high         -0.01197144  0.00707917  -1.691                0.09084 .
x9x9low          -0.00141812  0.00383769  -0.370                0.71174
x9x9medium                NA          NA      NA                    NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2227 on 14980 degrees of freedom
Multiple R-squared:  0.1989,    Adjusted R-squared:  0.1979
F-statistic: 206.6 on 18 and 14980 DF,  p-value: < 0.00000000000000022

> VIF(res)
[1] 1.248299
> bptest(res)

        studentized Breusch-Pagan test

data:  res
BP = 2527.2, df = 18, p-value < 0.00000000000000022
```

```
> res2 = lm(log(y)~x1 + x2 + x3 + x4 + x5 + x6+ x7+x8+x9) # Selected
> summary(res2)

Call:
lm(formula = log(y) ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 +
    x9)

Residuals:
     Min       1Q   Median       3Q      Max
-1.91447 -0.18688  0.04124  0.31185  1.18883

Coefficients: (2 not defined because of singularities)
                   Estimate  Std. Error t value             Pr(>|t|)
(Intercept)     -0.10576312  0.02469624  -4.283              0.00001859 ***
x1               0.42677700  0.02652117  16.092 < 0.0000000000000002 ***
x2              -0.15174422  0.00384414 -39.474 < 0.0000000000000002 ***
x3              -0.00029669  0.00009372  -3.166               0.00155 **
x4              -0.01340466  0.00296056  -4.528              0.00000601 ***
x5              -0.00252112  0.01189985  -0.212               0.83222
x6              -0.56762997  0.01012607 -56.056 < 0.0000000000000002 ***
x7               0.03246956  0.02914676   1.114               0.26530
x8x8accounting  -0.04651832  0.02068593  -2.249               0.02454 *
x8x8hr          -0.01620476  0.02100656  -0.771               0.44047
x8x8IT          -0.00201284  0.01740180  -0.116               0.90792
x8x8management   0.00376676  0.02307686   0.163               0.87034
x8x8marketing    0.00266689  0.01985647   0.134               0.89316
x8x8product_mng  0.01563664  0.01944205   0.804               0.42125
x8x8RandD       -0.02159977  0.02051011  -1.053               0.29230
x8x8sales        0.00636106  0.01250409   0.509               0.61096
x8x8support      0.01601610  0.01445450   1.108               0.26786
x8x8technical           NA          NA      NA                   NA
x9x9high        -0.02303106  0.01607877  -1.432               0.15205
x9x9low         -0.00356916  0.00871647  -0.409               0.68220
x9x9medium              NA          NA      NA                   NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5057 on 14980 degrees of freedom
Multiple R-squared:  0.2741,    Adjusted R-squared:  0.2732
F-statistic: 314.2 on 18 and 14980 DF,  p-value: < 0.00000000000000022

> VIF(res2)
[1] 1.377542
> bptest(res2)

        studentized Breusch-Pagan test

data:  res2
BP = 4067, df = 18, p-value < 0.00000000000000022
```

```
> x1_sq = x1^2
> res7 = lm(log(y)~x1 + x1_sq + x2 + x3 + x4 + x5 + x6+ x7+x8+x9) # Selected
> summary(res7)

Call:
lm(formula = log(y) ~ x1 + x1_sq + x2 + x3 + x4 + x5 + x6 + x7 +
    x8 + x9)

Residuals:
     Min       1Q   Median       3Q      Max
-1.91648 -0.18900  0.04211  0.31337  1.19177

Coefficients: (2 not defined because of singularities)
                  Estimate  Std. Error t value             Pr(>|t|)
(Intercept)      0.07733915  0.08344106   0.927              0.35401
x1              -0.12207398  0.24037981  -0.508              0.61157
x1_sq            0.38308673  0.16675617   2.297              0.02162 *
x2              -0.15131445  0.00384814 -39.321 < 0.0000000000000002 ***
x3              -0.00028206  0.00009392  -3.003              0.00268 **
x4              -0.01385183  0.00296653  -4.669           0.00000305 ***
x5              -0.00272106  0.01189847  -0.229              0.81911
x6              -0.57242669  0.01033769 -55.373 < 0.0000000000000002 ***
x7               0.03183736  0.02914390   1.092              0.27467
x8x8accounting  -0.04618029  0.02068350  -2.233              0.02558 *
x8x8hr          -0.01605582  0.02100366  -0.764              0.44462
x8x8IT          -0.00095965  0.01740536  -0.055              0.95603
x8x8management   0.00518470  0.02308182   0.225              0.82228
x8x8marketing    0.00279516  0.01985372   0.141              0.88804
x8x8product_mng  0.01492351  0.01944176   0.768              0.44274
x8x8RandD       -0.02095865  0.02050908  -1.022              0.30683
x8x8sales        0.00668661  0.01250311   0.535              0.59280
x8x8support      0.01631458  0.01445302   1.129              0.25900
x8x8technical            NA          NA      NA                  NA
x9x9high        -0.02219271  0.01608062  -1.380              0.16758
x9x9low         -0.00344311  0.00871540  -0.395              0.69280
x9x9medium               NA          NA      NA                  NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5057 on 14979 degrees of freedom
Multiple R-squared:  0.2743,    Adjusted R-squared:  0.2734
F-statistic:   298 on 19 and 14979 DF,  p-value: < 0.00000000000000022

> VIF(res7)
[1] 1.378027
> bptest(res7)

        studentized Breusch-Pagan test

data:  res7
BP = 4027.1, df = 19, p-value < 0.00000000000000022
```

```
> x3_sq = x3^2
> res8 = lm(log(y)~x1 + x1_sq + x2 + x3 + x3_sq + x4 + x5 + x6+ x7+x8+x9)
> summary(res8)

Call:
lm(formula = log(y) ~ x1 + x1_sq + x2 + x3 + x3_sq + x4 + x5 +
    x6 + x7 + x8 + x9)

Residuals:
     Min       1Q   Median       3Q      Max
-1.95267 -0.20856  0.03769  0.30999  1.50707

Coefficients: (2 not defined because of singularities)
                   Estimate  Std. Error t value             Pr(>|t|)
(Intercept)     -2.04955647  0.10696904 -19.160 < 0.0000000000000002 ***
x1              -0.30921261  0.23336536  -1.325              0.185186
x1_sq            0.47197981  0.16186025   2.916              0.003551 **
x2              -0.13642945  0.00376645 -36.222 < 0.0000000000000002 ***
x3               0.02223220  0.00074544  29.824 < 0.0000000000000002 ***
x3_sq           -0.00005599  0.00000184 -30.431 < 0.0000000000000002 ***
x4              -0.01046697  0.00288112  -3.633              0.000281 ***
x5              -0.00346605  0.01154728  -0.300              0.764058
x6              -0.48096803  0.01047304 -45.924 < 0.0000000000000002 ***
x7               0.03855868  0.02828450   1.363              0.172826
x8x8accounting  -0.04431378  0.02007307  -2.208              0.027286 *
x8x8hr          -0.01773944  0.02038375  -0.870              0.384165
x8x8IT           0.00335779  0.01689218   0.199              0.842439
x8x8management  -0.00659873  0.02240384  -0.295              0.768353
x8x8marketing   -0.00075995  0.01926803  -0.039              0.968539
x8x8product_mng  0.01578078  0.01886790   0.836              0.402953
x8x8RandD       -0.01918551  0.01990378  -0.964              0.335105
x8x8sales        0.00223801  0.01213492   0.184              0.853681
x8x8support      0.01504347  0.01402646   1.073              0.283510
x8x8technical           NA          NA      NA                    NA
x9x9high        -0.01742844  0.01560674  -1.117              0.264130
x9x9low         -0.00308279  0.00845815  -0.364              0.715507
x9x9medium              NA          NA      NA                    NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4907 on 14978 degrees of freedom
Multiple R-squared:  0.3166,    Adjusted R-squared:  0.3157
F-statistic: 346.9 on 20 and 14978 DF,  p-value: < 0.00000000000000022

> VIF(res8)
[1] 1.463227
> bptest(res8)

        studentized Breusch-Pagan test

data:  res8
BP = 2925.9, df = 20, p-value < 0.00000000000000022
```

```
> x4_sq = x4^2
> res11 = lm(log(y)~x1 + x1_sq + x2 + x3 + x3_sq + x4+ x4_sq + x5 + x6+ x7+x8+x9)
> summary(res11)

Call:
lm(formula = log(y) ~ x1 + x1_sq + x2 + x3 + x3_sq + x4 + x4_sq +
    x5 + x6 + x7 + x8 + x9)

Residuals:
     Min       1Q   Median       3Q      Max
-1.94062 -0.21131  0.03526  0.30906  1.51273

Coefficients: (2 not defined because of singularities)
                    Estimate   Std. Error t value             Pr(>|t|)
(Intercept)     -1.842427659  0.109668550 -16.800 < 0.0000000000000002 ***
x1              -0.403897791  0.233133825  -1.732             0.083210 .
x1_sq            0.542190707  0.161727898   3.352             0.000803 ***
x2              -0.131893766  0.003798432 -34.723 < 0.0000000000000002 ***
x3               0.022062160  0.000744077  29.650 < 0.0000000000000002 ***
x3_sq           -0.000055531  0.000001837 -30.232 < 0.0000000000000002 ***
x4              -0.100262137  0.011299743  -8.873 < 0.0000000000000002 ***
x4_sq            0.009080638  0.001105101   8.217 0.000000000000000226 ***
x5              -0.004933732  0.011523107  -0.428             0.668541
x6              -0.458718429  0.010794976 -42.494 < 0.0000000000000002 ***
x7               0.037138821  0.028222430   1.316             0.188217
x8x8accounting  -0.044548089  0.020028661  -2.224             0.026149 *
x8x8hr          -0.016942187  0.020338869  -0.833             0.404861
x8x8IT           0.002618570  0.016855037   0.155             0.876541
x8x8management  -0.023188914  0.022445248  -1.033             0.301559
x8x8marketing   -0.001888427  0.019225880  -0.098             0.921756
x8x8product_mng  0.014618751  0.018826672   0.776             0.437471
x8x8RandD       -0.016544662  0.019862329  -0.833             0.404877
x8x8sales        0.000100809  0.012110861   0.008             0.993359
x8x8support      0.013235162  0.013997149   0.946             0.344388
x8x8technical             NA           NA      NA                   NA
x9x9high        -0.022252972  0.015583260  -1.428             0.153311
x9x9low         -0.002565664  0.008439660  -0.304             0.761131
x9x9medium                NA           NA      NA                   NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4897 on 14977 degrees of freedom
Multiple R-squared:  0.3196,    Adjusted R-squared:  0.3187
F-statistic: 335.1 on 21 and 14977 DF,  p-value: < 0.00000000000000022

> VIF(res11)
[1] 1.469823
> bptest(res11)

        studentized Breusch-Pagan test

data:  res11
BP = 3398.8, df = 21, p-value < 0.00000000000000022
```

Model Selected:

Ln_Satisfaction_level = -1.84 + (-0.4038) * last_evaluation + (0.542) * last_evaluation_sq

+ (-0.132) * number_project

+ (0.022) * average_montly_hours + (-0.000055) * average_montly_hours_sq

+ (-0.1) time_spend_company + (0.009) time_spend_company_sq

+ (-0.0049) * Work_accident

+ (-0.4587) * left

+ (0.0371) * promotion_last_5years

+(-0.044) salesxaccounting

+(-0.0169) salesxhr

+ (0.090026)salesxIT

+ (-0.0231)salesxmanagement

+ (-0.0018) salesxmarketing

+ (0.0146) salesxproduct_mng

+ (-0.0165)salesxRandD

+ (0.0001) salesxsales

+ (0.0132) salesxsupport

+ (-0.0222) * salaryxhigh + (-0.0025) * salaryxlow

Base Group: Salary: Medium, Dept : Technical

Inference:

- If there is 1-unit increase in last evaluation, then the satisfaction decreases by 40.38% till 1.49 and then starts increasing. Keeping everything else constant.
- If the no of projects increases by 1 unit then the satisfaction decreases by 13.2% for an employee with medium salary and is in the Technical Dept. Keeping everything else constant.
- If the average monthly hours increase by 1 the satisfaction increases by 2.2% till the turning point of 80 after that it starts decreasing. Keeping everything else constant
- If the employee has left the company, then the overall satisfaction decreases by 45.87% for an employee who used to have medium salary and was in the Technical Dept. Keeping everything else constant.
- If the time spent in the company increase by 1 the satisfaction decreases by 10% till the turning point of 22.22 after that it starts increasing. Keeping everything else constant

Model Selection:

We regressed the independent variables: last_evaluation, number_project, average_montly_hours, time_spend_company, Work_accident, left, promotion_last_5years, sales, salary against log satisfaction. To get the percentage change in the satisfaction level. We expected to have diminishing returns in last_evaluation, average_monthly_hours and time_spend_company. We got an R^2 of 31.96%.

Conclusion

Job satisfaction as we saw previously is a very important factor that affects the turnover of employees in an organization. There are several parameters that affect it. From our results, we found that there are top four important factors affecting the job satisfaction: last_evaluation, number of projects, average monthly hours and time spent in the company. From these, last_evaluation that is, the last time an employee's performance was evaluated came out to be the most significant factor of all. This means that if an employee is given a better rating, but if he was not given a good rating, his satisfaction level decreased, and this led to him leaving the company. Keeping everything else constant, if there is 1-unit increase in last_evaluation, then the satisfaction decreases by 40.38% but it decreases till 1.49 and then starts increasing. This could be because the evaluation was long back, and they forgot about it.

Finally, it is important to note that satisfaction level of an is the most important factor to retain a skilled employee. For this, if the organization considers factors like the last evaluation, promotion, average number of hours they make an employee spend per month etc., they could reduce attrition.

References

Clark, A. E. (2001). “What really matters in a job? Hedonic measurement using quit data.” Labour Economics, Vol. 8, 223-242.

Shields, M. and Ward, M. (2001), “Improving nurse retention in the National Health Service in Enland: the impact of job satisfaction on intentions to quit.” Journal of Health Economics, Vol. 20, 677-701

Brayfield & Crockett, 1955 Brayfield, A. H., & Crockett, W. H. (1955). Employee attitudes and employee performance. Psychological Bulletin, 52, 396–424.10.1037/h0045899

Goslin, 2005 Goslin, P. A. (2005). Managing employee satisfaction of volunteers in South African sport. African Journal for Physical Health Education, 12, 30–40.

Chan et al., 2000 Chan, K. C., Gee, M. V., & Steiner, T. L. (2000). Employee happiness and corporate financial performance. Financial Practice and Education, 10, 47–52.

Mohr & Puck, 2007 Mohr, A. T., & Puck, J. F. (2007). Role conflict, general manager job satisfaction and stress and the performance of international joint ventures. European Management Journal, 25, 25–35.

https://pdfs.semanticscholar.org/19a4/3135ef0f3b95d0853fd1d33a056929e23b12.pdf