Predicting Employee Churn

by

Celestine Chidi Azonobi

Objectives:

- Analyze employee churn.
- Find out why employees are leaving the company
- Predict who will leave the company.

What is Employee Churn?

Employee churn can be defined as a leak or departure of an intellectual asset from a company or organization.

In simple words, we can put it as when employees leave an organization is known.

The Following points will help us to understand this in a better way:

- Business chooses the employee or hire someone.
- Employees will be the face of your company, and collectively, the employees produce everything your company does.
- Employee churn put companies or organizations in difficult situations because it requires time and effort in finding and training a replacement.

I’ll be using the next few slides to show how I analysed a company’s data whose employees churned as well as my methodology.

Basic information about our data

There are two sets of data:

- Existing Employees.
- Employees who left.

- I used Python's pandas for the data analysis.
- Both data contain an initial 1,428 rows(for the Existing Employees dataset) and 3571 rows respectively , with 10 columns each.
- I combined both datasets for proper analysis.
- I reserved from data from the existing employees dataset to be used later in testing the machine algorithm.
- Both Cluster analysis and Classification prediction were carried out.

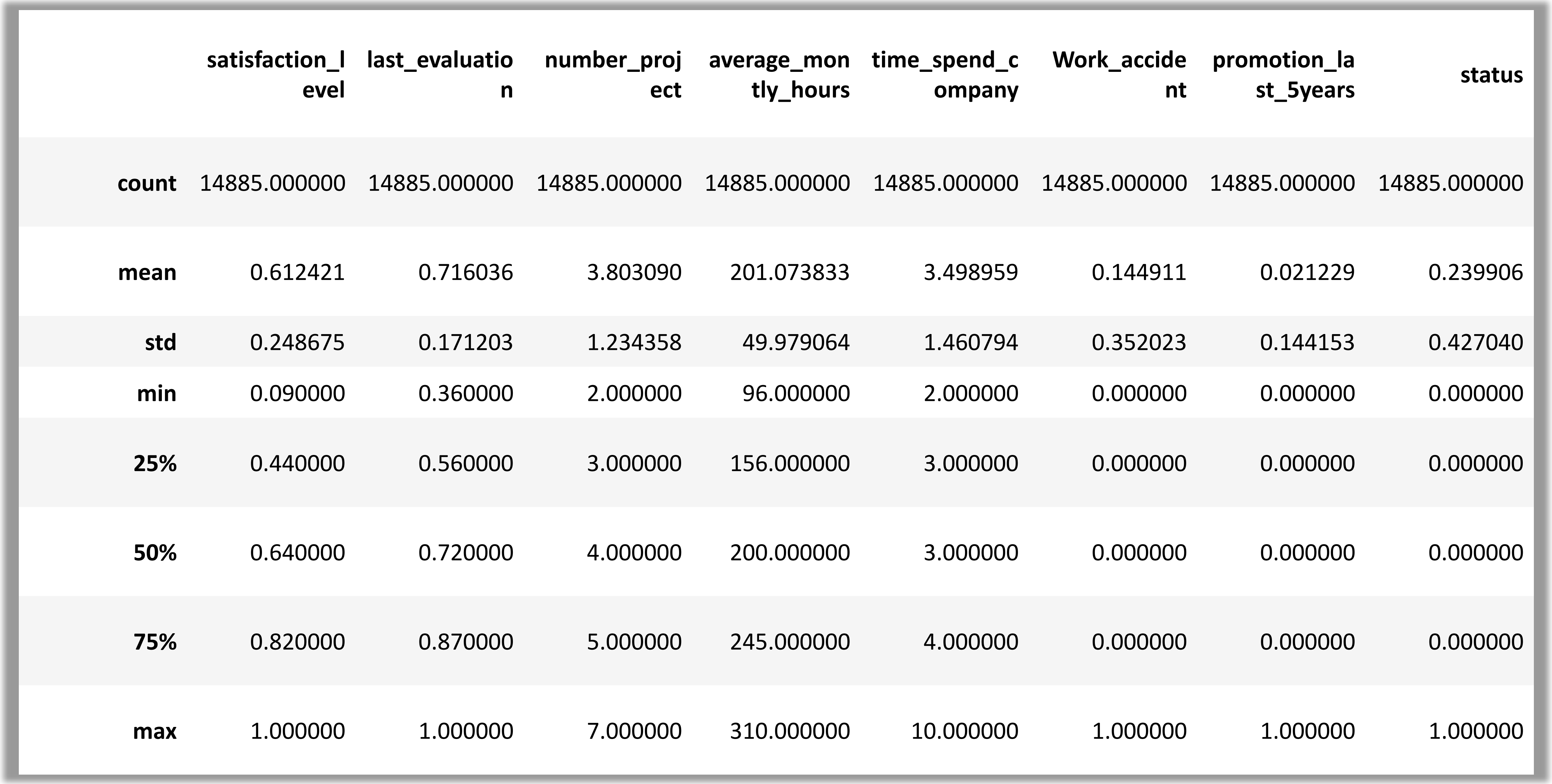
We can describe 10 columns in detail as:

- **satisfaction_level:** The Employees’ satisfaction point, which ranges from 0-1.
- **last_evaluation:** Evaluated performance by the employer, which also ranges from 0-1.
- **number_projects:** How many numbers of projects assigned to an employee?
- **average_monthly_hours:** Average numbers of hours worked by an employee in a month.
- **time_spent_company:** Simply means employee experience. The number of years spent by an employee in the company.
- **work_accident:** Whether an employee has had a work accident or not.
- **promotion_last_5years:** Whether an employee has had a promotion in the last 5 years or not.
- **depts:** Employee's working department/division.
- **salary:** Salary level of the employee such as low, medium and high.
- **Status(created during analysis):** Whether the employee has left the company or not.

0 – Existing Employees.

1- Employee who had left.

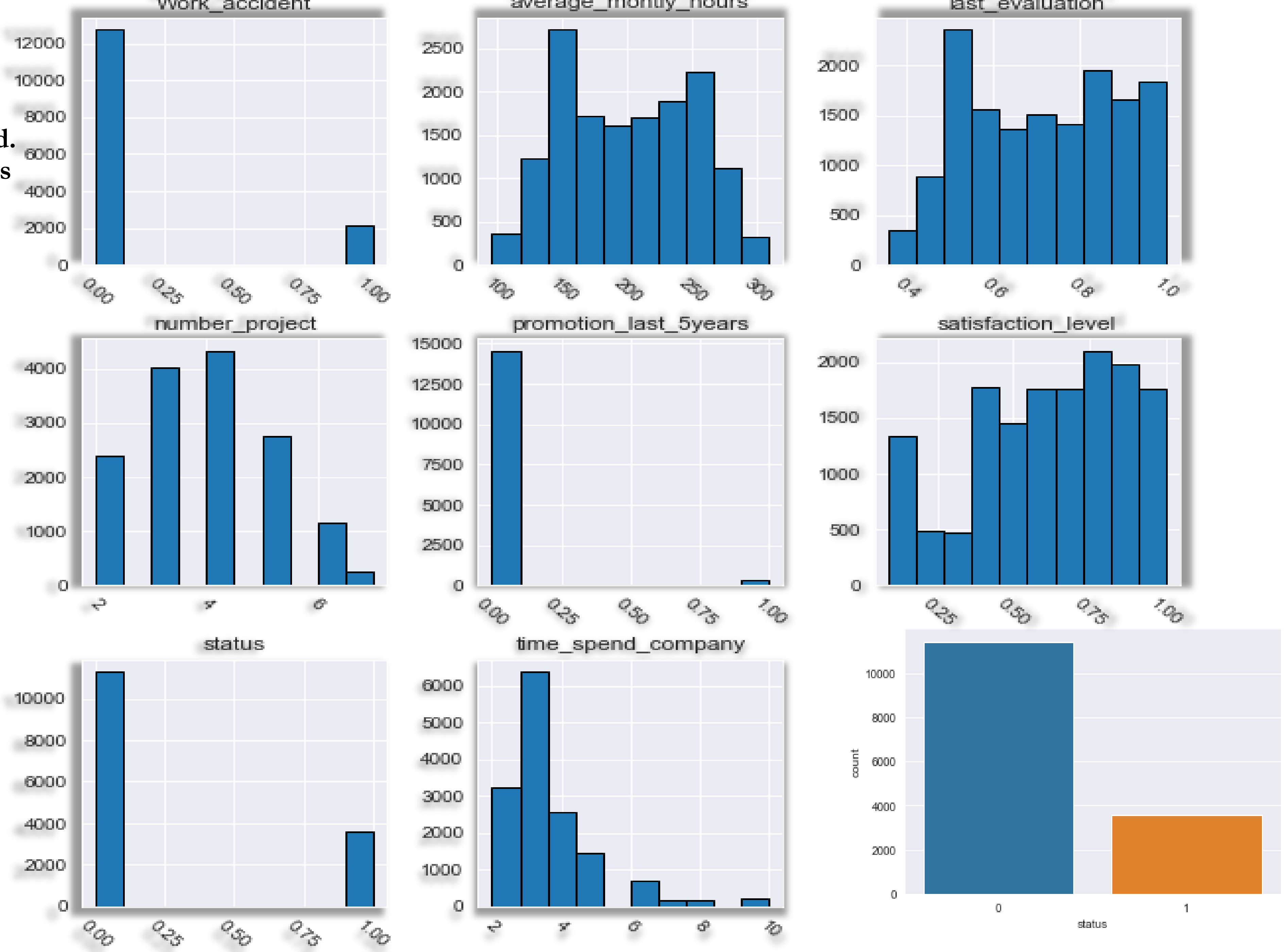
Below is the summary statistics of the combined dataset.



	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	promotion_last_5years	status
count	14885.000000	14885.000000	14885.000000	14885.000000	14885.000000	14885.000000	14885.000000	14885.000000
mean	0.612421	0.716036	3.803090	201.073833	3.498959	0.144911	0.021229	0.239906
std	0.248675	0.171203	1.234358	49.979064	1.460794	0.352023	0.144153	0.427040
min	0.090000	0.360000	2.000000	96.000000	2.000000	0.000000	0.000000	0.000000
25%	0.440000	0.560000	3.000000	156.000000	3.000000	0.000000	0.000000	0.000000
50%	0.640000	0.720000	4.000000	200.000000	3.000000	0.000000	0.000000	0.000000
75%	0.820000	0.870000	5.000000	245.000000	4.000000	0.000000	0.000000	0.000000
max	1.000000	1.000000	7.000000	310.000000	10.000000	1.000000	1.000000	1.000000

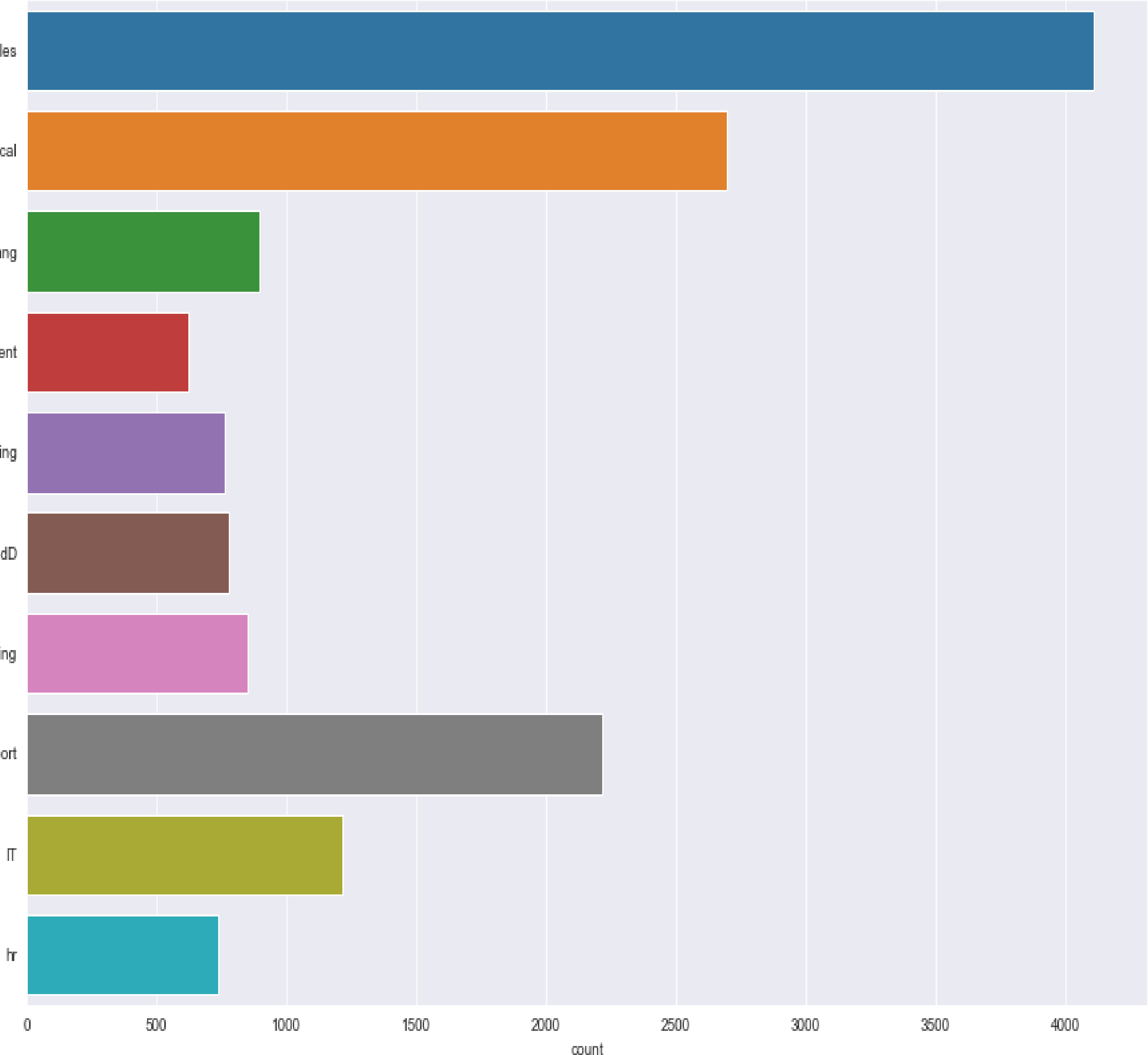
Distribution of numerical columns

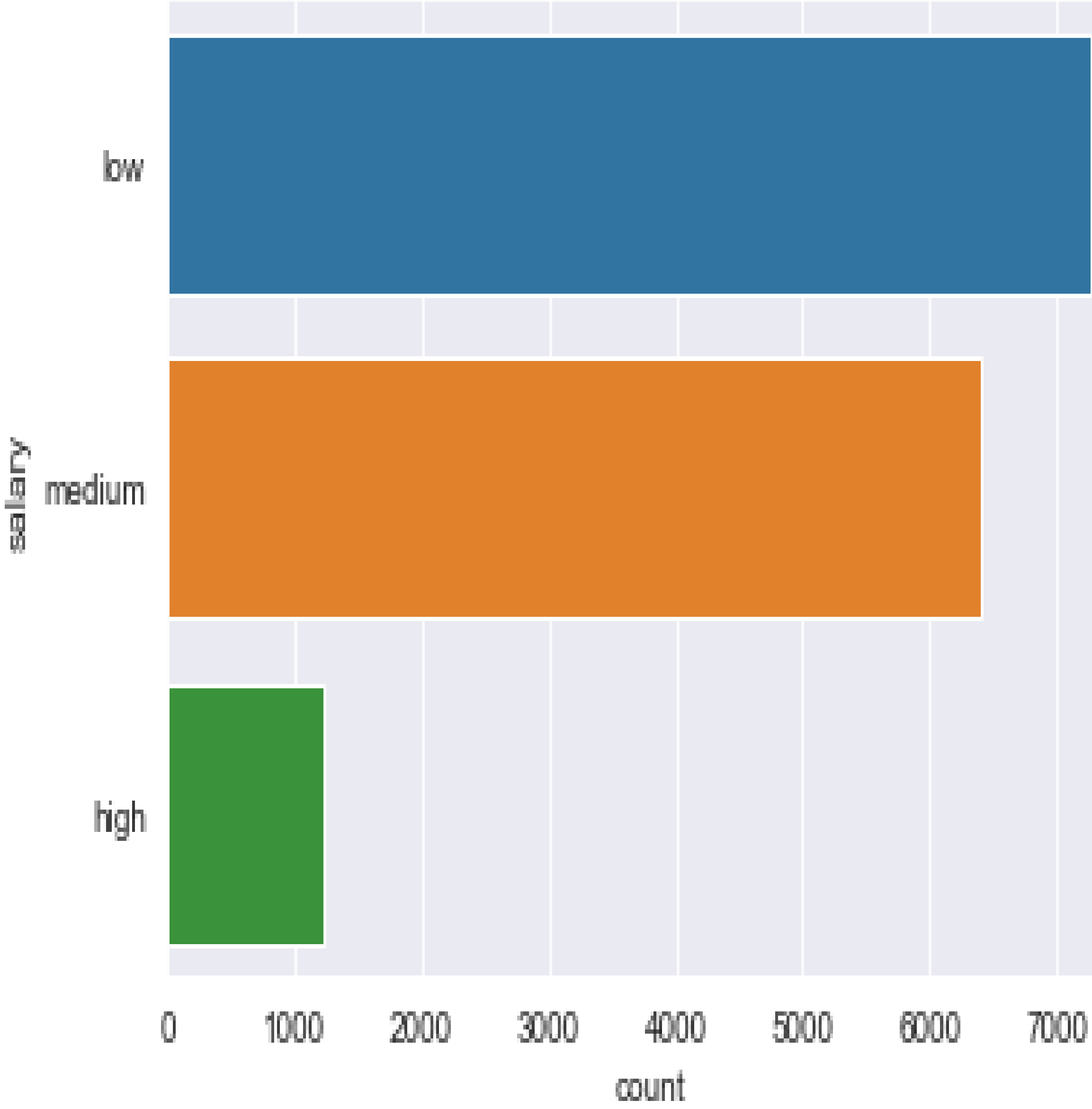
- 3,571 left, and 11,314 stayed.
- 24% of the total employees left the company.



Analysing categorical features.

Sales dept had the most no. of employees.
Majority are low on salary.





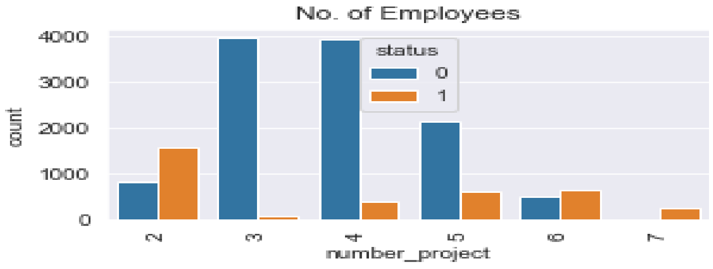
Joint plot of numerical columns

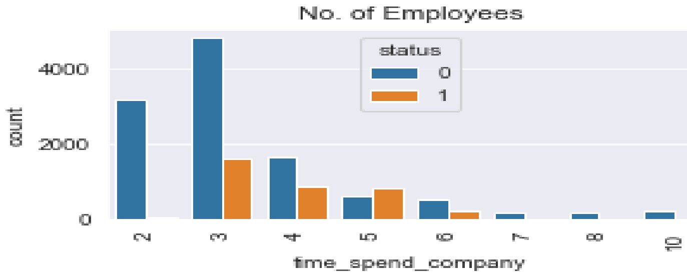
You can observe the following points in this visualization:

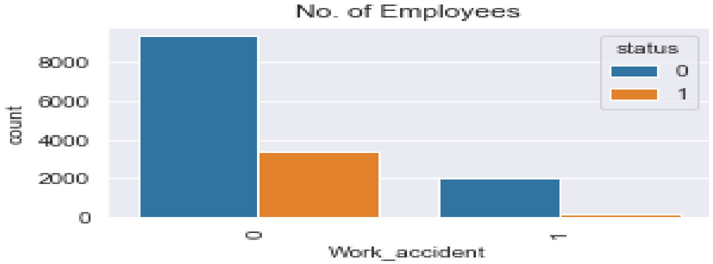
- Those employees who have the number of projects from 4 and above left the company it seems to like that they were overloaded with work.
- Employees with 3 to 5 years experience are leaving more. The ones with more experience are not leaving because of affection/affiliation with the company.
- Those who got promotion in last 5 years they didn't leave.
- More Employees left from the sales department.
- Employees with low salary left the department more.

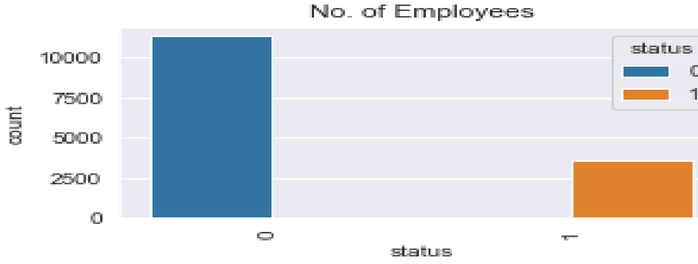
In summary, Those who left:

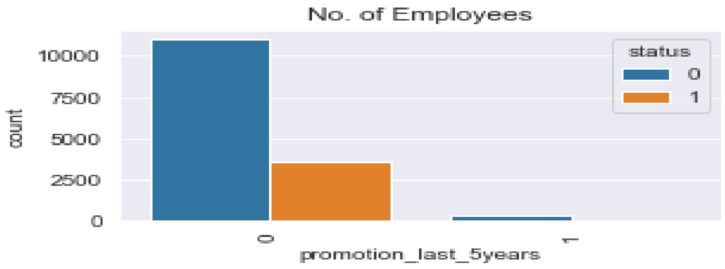
- had less satisfaction at the company
- spent the most hours at work in a month
- spent more time at the company
- had lower number of accidents
- had less promotion in the last 5 years.
- People with *3 years* of experience tend to leave often

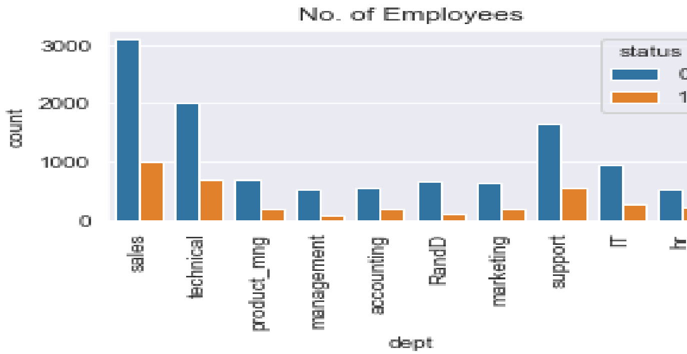


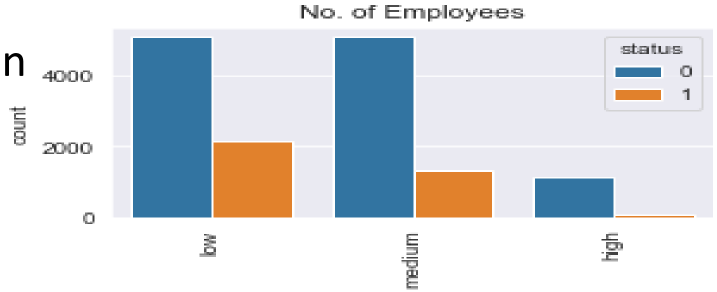






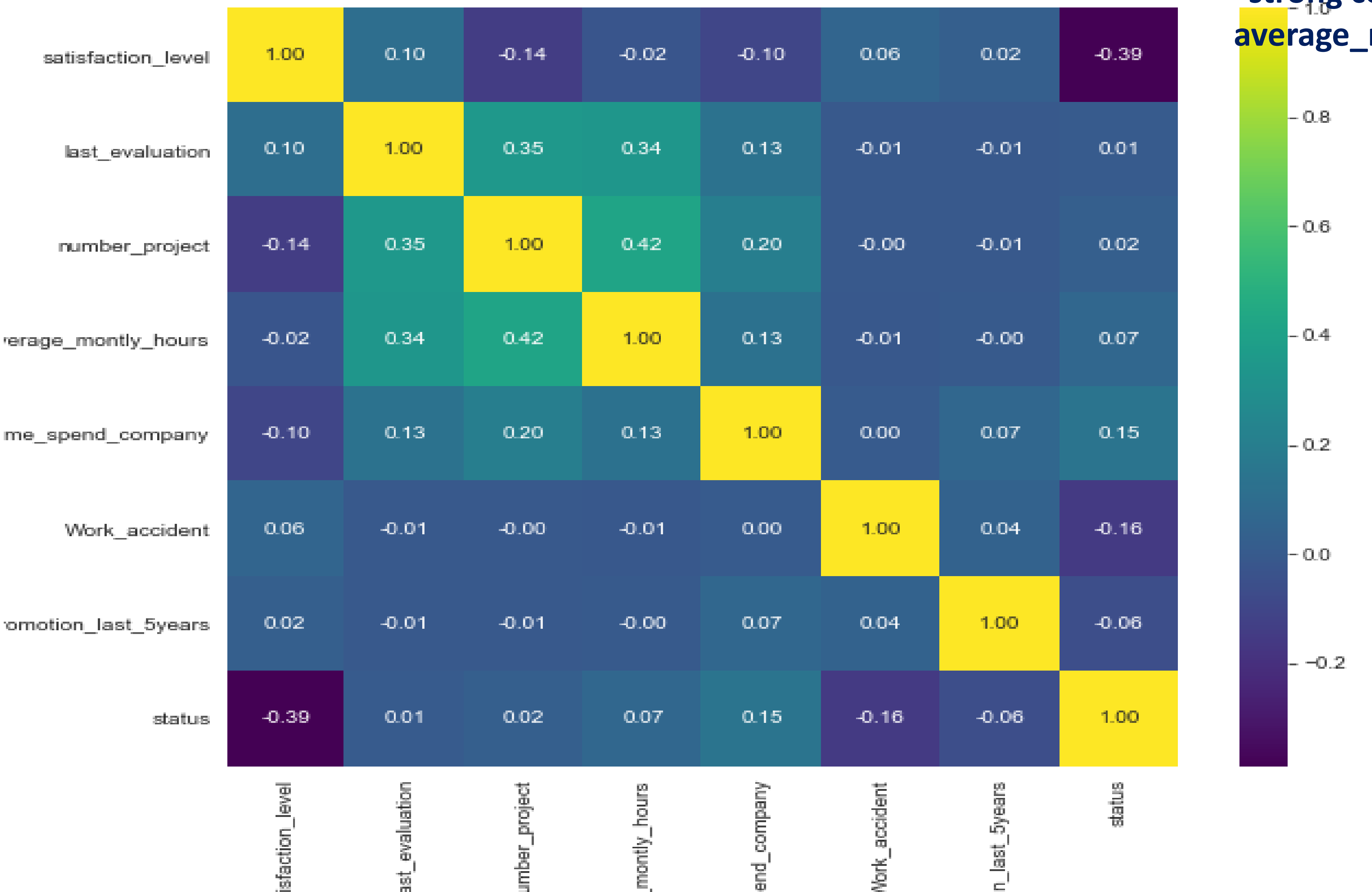




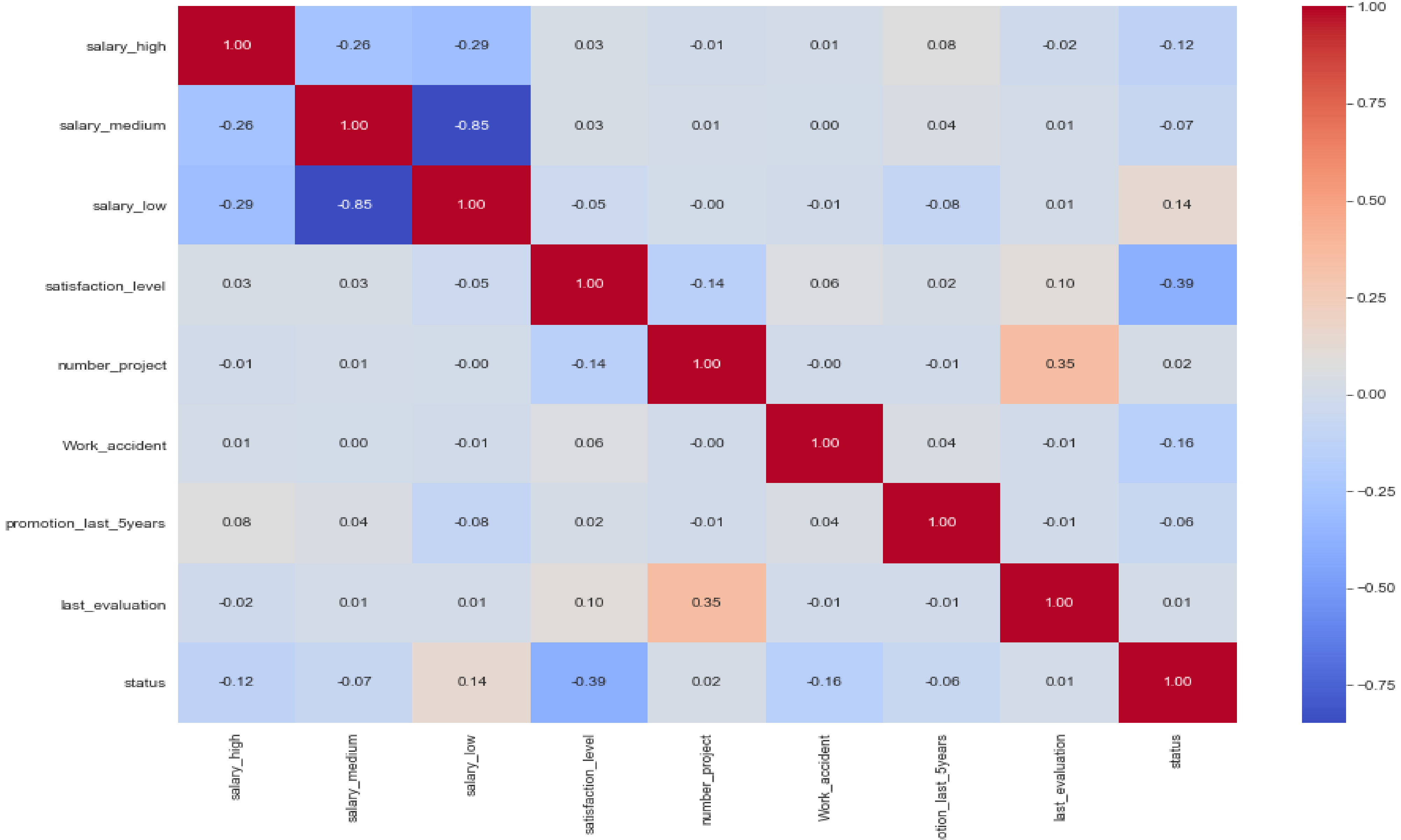


A heatmap showing different correlations of each numerical column.

- -Stong correlation btw Time_spent and
- -strong corr. btw last_evaluation and average_monthly hrs, number_project



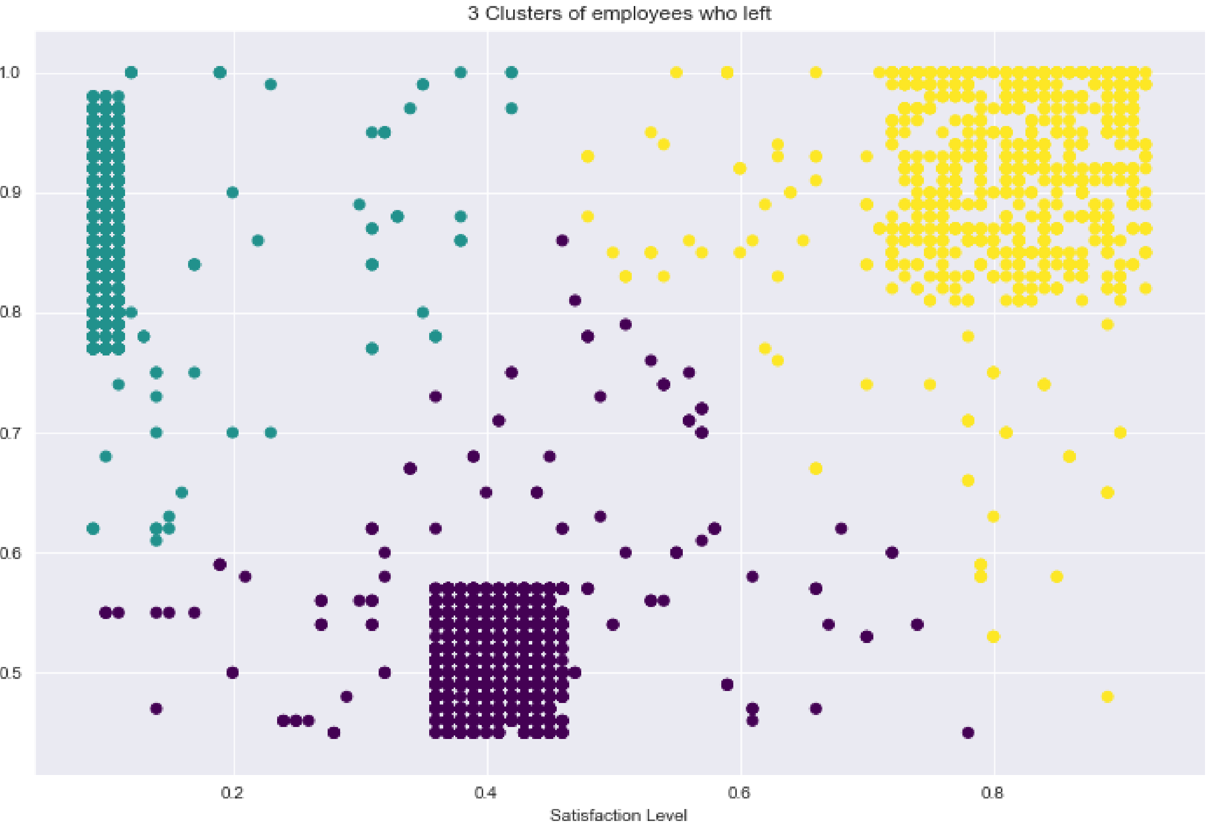
This heatmap shows correlation btw salary and status.

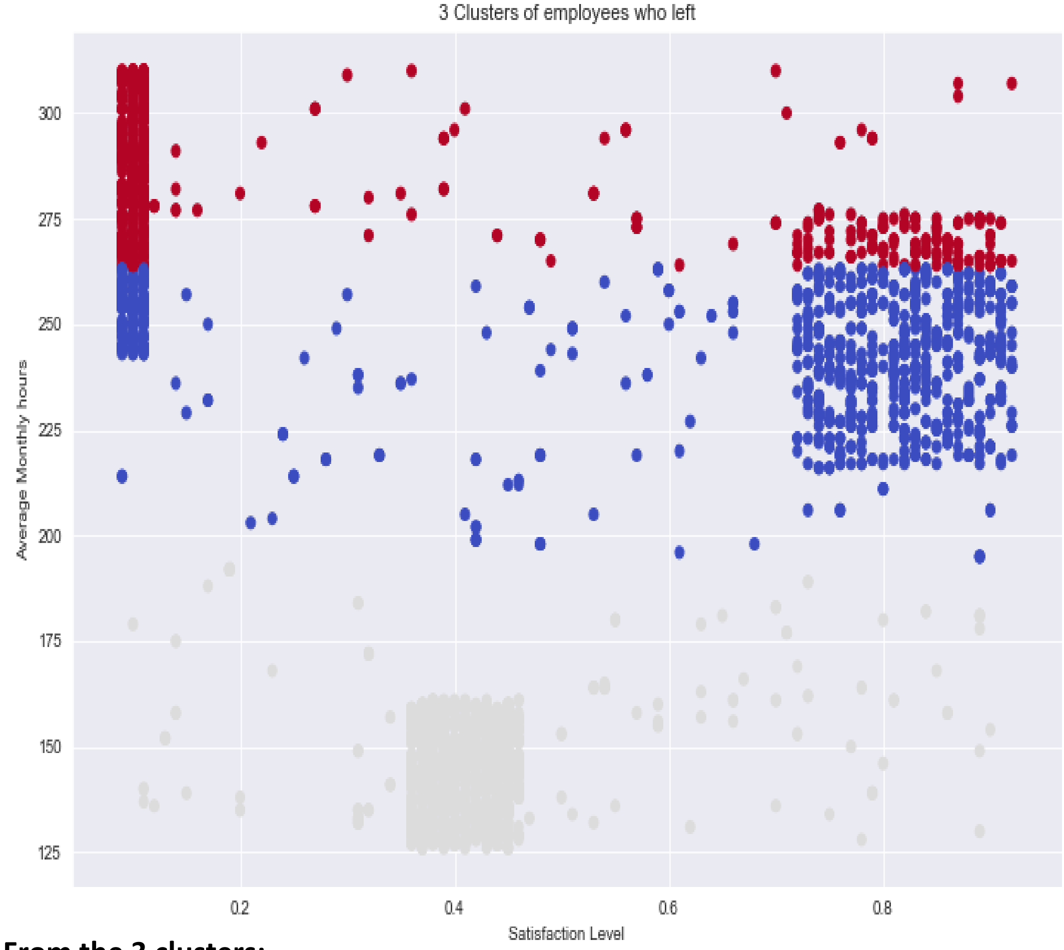


	salary_high	salary_medium	salary_low	satisfaction_level	number_project	Work_accident	otion_last_5years	last_evaluation	status
salary_high	1.00	-0.26	-0.29	0.03	-0.01	0.01	0.08	-0.02	-0.12
salary_medium	-0.26	1.00	-0.85	0.03	0.01	0.00	0.04	0.01	-0.07
salary_low	-0.29	-0.85	1.00	-0.05	-0.00	-0.01	-0.08	0.01	0.14
satisfaction_level	0.03	0.03	-0.05	1.00	-0.14	0.06	0.02	0.10	-0.39
number_project	-0.01	0.01	-0.00	-0.14	1.00	-0.00	-0.01	0.35	0.02
Work_accident	0.01	0.00	-0.01	0.06	-0.00	1.00	0.04	-0.01	-0.16
promotion_last_5years	0.08	0.04	-0.08	0.02	-0.01	0.04	1.00	-0.01	-0.06
last_evaluation	-0.02	0.01	0.01	0.10	0.35	-0.01	-0.01	1.00	0.01
status	-0.12	-0.07	0.14	-0.39	0.02	-0.16	-0.06	0.01	1.00

Cluster Analysis: I used the k-means cluster method with cluster of 3.

Employees who left:





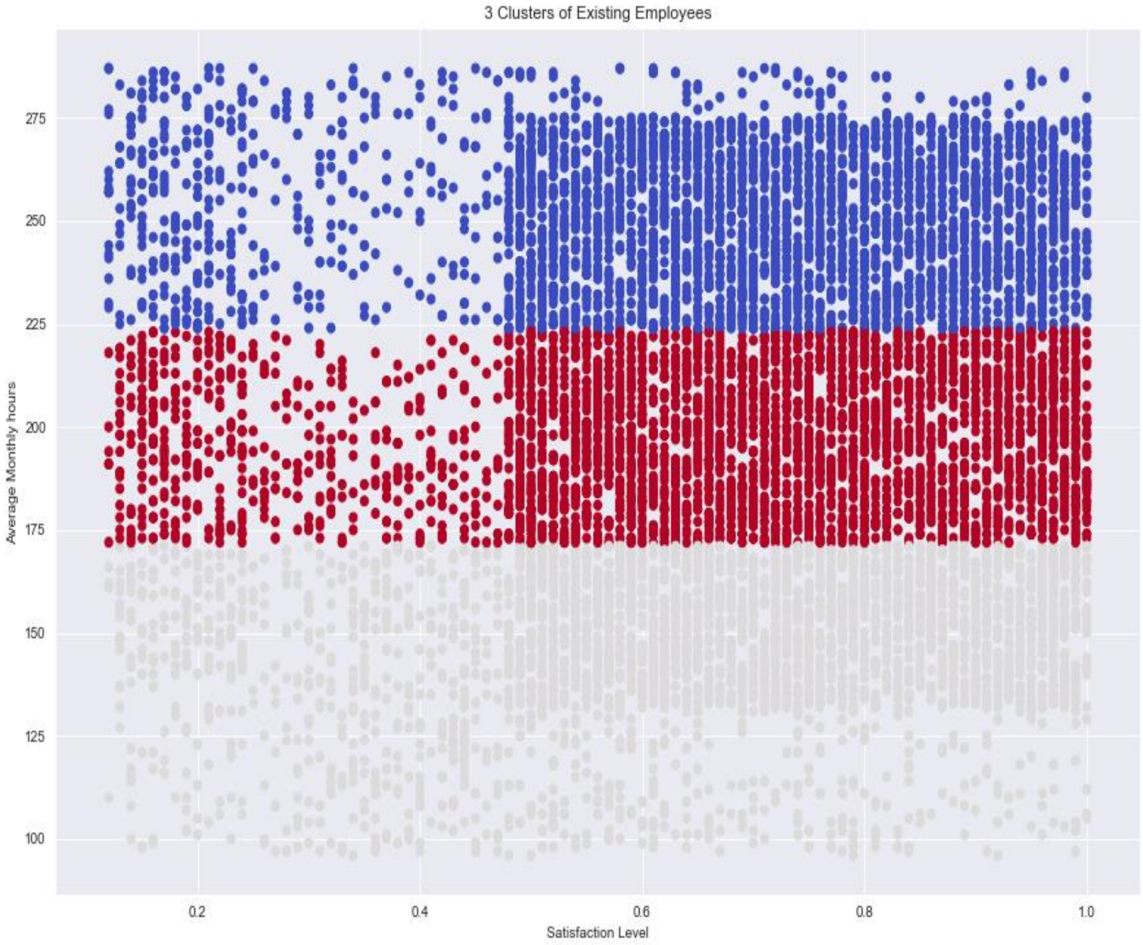
From the 3 clusters:
y - had high evaluation and high satisfaction rate...
These ones must have left for due to salary and promotion.
g - had high evaluation but low satisfaction-
(They definitely thought about leaving because of low satisfaction)
p = Relatively low evaluation and low satisfaction rate.
Hence the reason they left.

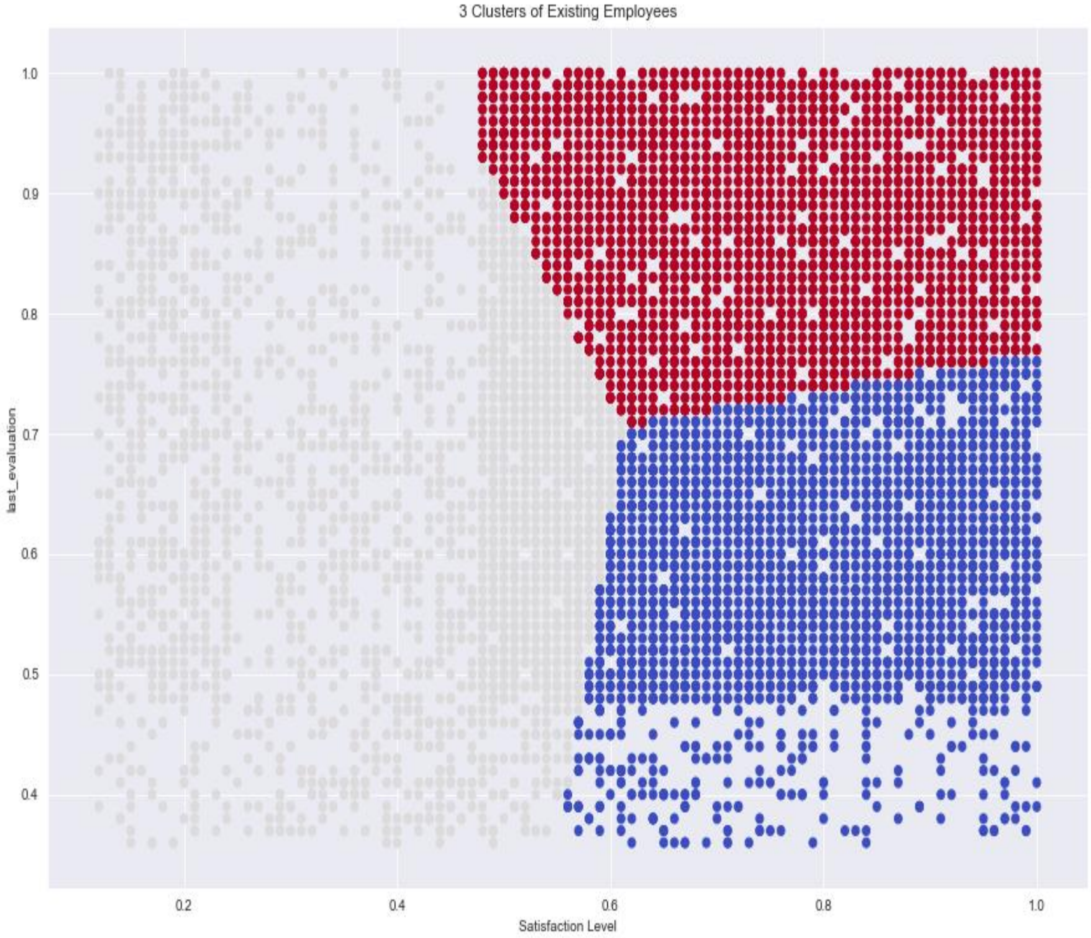
From the 3 clusters:

- **m = more time at the office with low satisfaction rate. Few spent much time with high satisfaction rate...**
  it seems they had a higher promotion rate, hence they were satisfied.
- **b = spent an average amount of time with mostly high satisfaction rate...**
- **g = Relatively spent lower time with low satisfaction rate. Left due to lack of promotion and/or low salary.**

Cluster Analysis: I used the k-means cluster method with cluster of 3.

Employees who left:





So those that may likely leave the company in future:

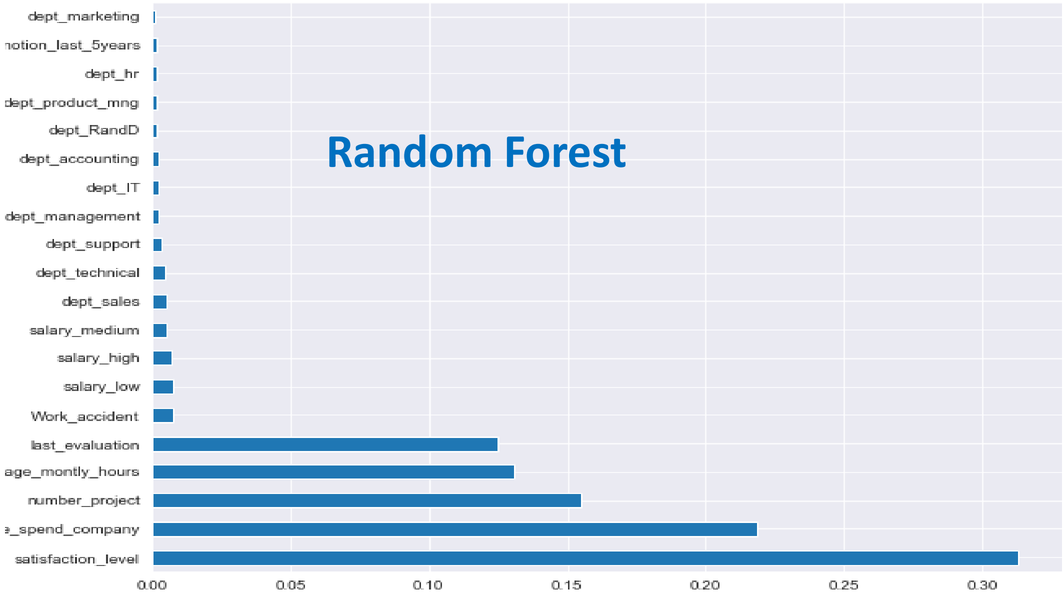
b - Those who spent a lot of hours(>=225 hrs) but have a low satisfaction rate.

m - Those that spend a monthly average >=170 hrs and <=225 hrs with low satisfaction rate(Not that many though).

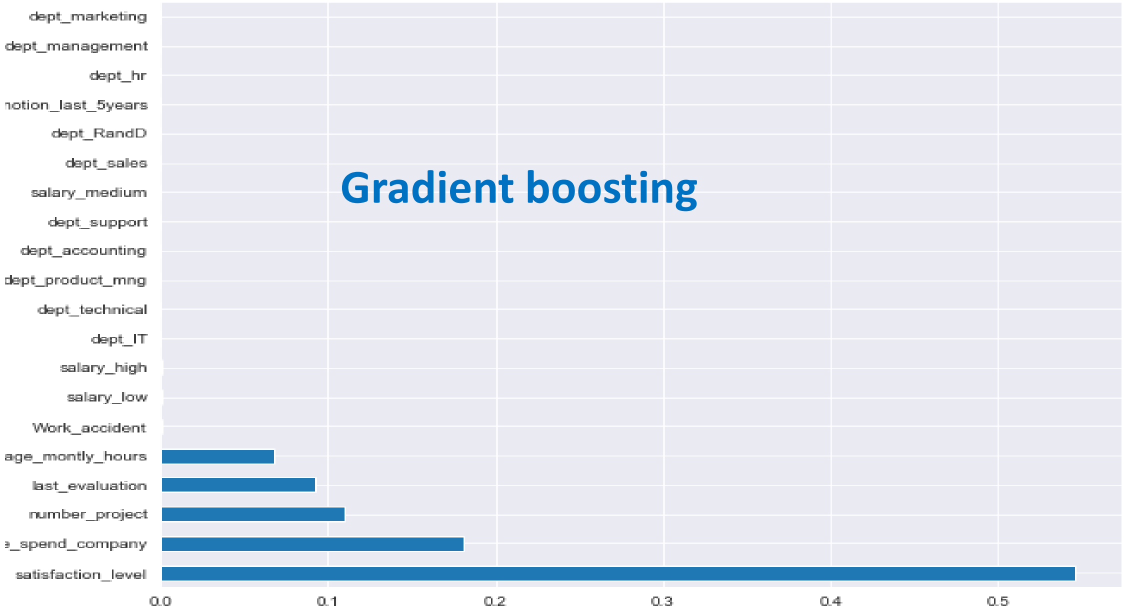
g - They spend a low monthly average <=170 hrs with high satisfaction rate. Most of them ones will be staying. The ones with very low satisfaction rate may also leave due to other factors.

In this cluster, we can conclude that based on evaluation and satisfaction, the cluster of employees with gray all have low satisfaction and may therefo likely leave in the near future.

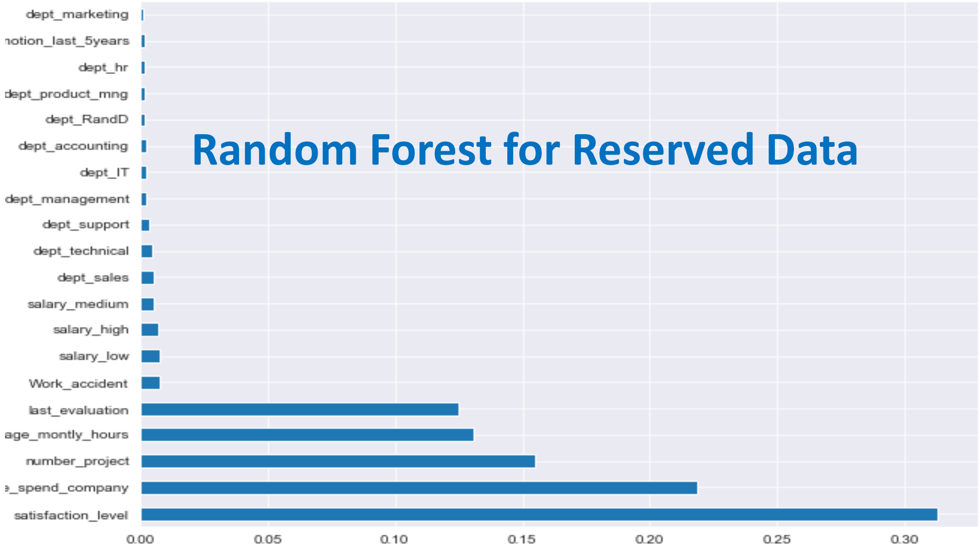
Predicting with Gradient boosting and Random Forest Classifiers.



Random Forest



Gradient boosting



Random Forest for Reserved Data

Model Prediction Accuracies:

Training Set Accuracy: 97.6(Gradient Boosting)
Test Set: 98%(Gradient Boosting)

Training Set Accuracy: 99.8% (Random Forest)
Test set: 99% (Random Forest)

Using the reserved Data, Random Forest Classifier Predicted that no employee would be leaving soon.

From the feature importance plots, we can certainly conclude that ***Employee satisfaction*** had the most important influence on the employee's decision to either leave or stay. Other influencing features were **time spent in the company, number of projects, last_evaluation and Average monthly hours**

Data Analysis and Visualization Summary:

From the analyses, the following features mostly influenced a person to leave the company:

- **Satisfaction_leve**l: Employees are far more likely to quit if their satisfaction level is low.
- **Time with Company:** Here, The three-year mark looks like a time to be a crucial point in an employee's career. Most of them quit their job around the three-year mark. Another important point is 6-years point, where the employee is very unlikely to leave.
- **Number Of Projects:** Employee engagement is another critical factor that influences the employee to leave the company. Employees with 3-5 projects are less likely to leave the company. The employee with less and more number of projects are likely to leave.
- **Average Monthly hours:** Employees who feel that they spend more hours working but not being appreciated enough by the company.
- **Last_Evaluation:** Employees feel he’s not been evaluated as regularly as they would like.

Hence they feel they’re been overlooked.

Others are low salary and work accident.

So the company should work on these aspects of the employees’ welfare.

Important Notice: The dataset is actually an imbalanced dataset since there are more samples of existing employees than employees who left. It can be improved by adding more samples.