



CAPSTONE PROJECT-2

---

# EMPLOYEE CHURN PREDICTION

---

Group-4

# CONTENTS

- Group Members
- Meeting Schedule
- Methodology
- EDA
- Modeling
  - K-Means
  - Gradient Boosting
  - KNN
  - Random Forest
  - Catboost
- Model Evaluation
- Model Deployment

# GROUP MEMBERS

C8124-JACK

C8125-MUSTAFA

C8250-ONUR

C8278-ENGİN

C8292-KEN

C8301-SAM

C8307-HAKAN

C8315-HALİT

C8399-BENJAMIN

C8492-SEMIH

C9231-HÜSEYİN



# MEETING SCHEDULE

10-13 December 2021 : Individual study

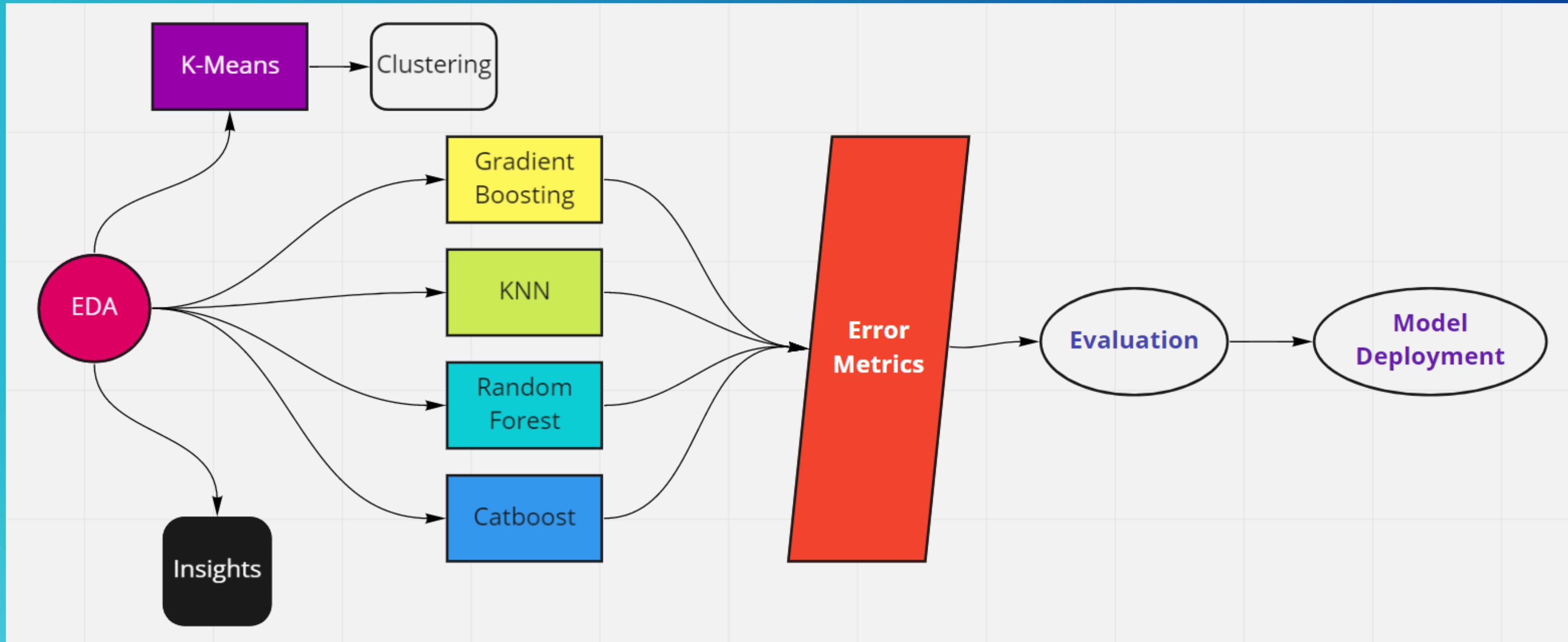
13 December 2021 : EDA & K-Means Modelling  
20:00 (IST)

14 December 2021 : Gradient Boosting & KNN & Random Forest & Catboost  
20:00 (IST)

15 December 2021 : Model Deployment  
20:00 (IST)

16 December 2021 : Review  
13:00 (IST)

# METHODOLOGY



# EDA (EXPLORATORY DATA ANALYSIS)

Columns : 10

Rows : 14.999

Duplicated: 3.008

Null: None

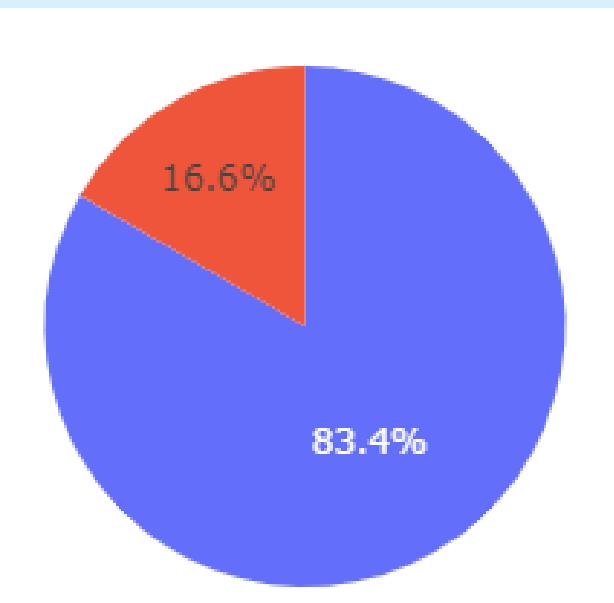
- **satisfaction\_level:** It is employee satisfaction point, which ranges from 0-1.
- **last\_evaluation:** It is evaluated performance by the employer, which also ranges from 0-1.
- **number\_projects:** How many of projects assigned to an employee?
- **average\_monthly\_hours:** How many hours in average an employee worked in a month?
- **time\_spent\_company:** time\_spent\_company 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.
- **Departments:** Employee's working department/division.
- **Salary:** Salary level of the employee such as low, medium and high.
- **left:** Whether the employee has left the company or not.

Number of Uniques:	
satisfaction_level	92
last_evaluation	65
number_project	6
average_montly_hours	215
time_spend_company	8
Work_accident	2
left	2
promotion_last_5years	2
Departments	10
salary	3

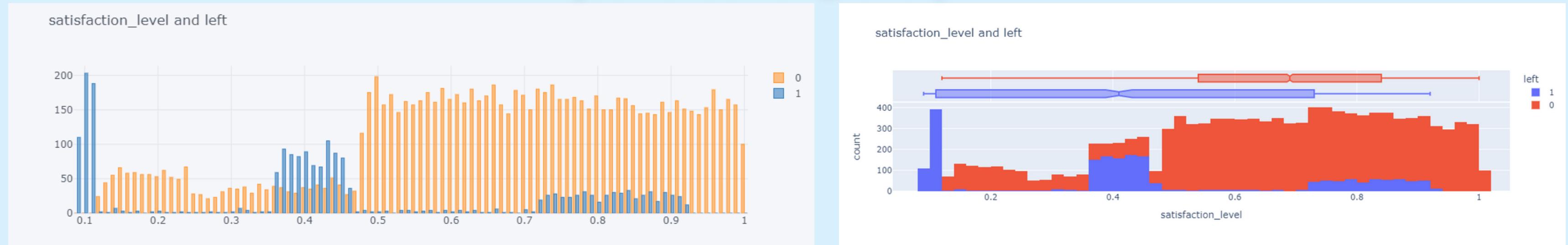
Percentage of left-1: % 16.6 --> (1991 observations for left-1)  
 Percentage of left-0: % 83.4 --> (10000 observations for left-0)

- 'left' column has binary type values.
- We have an imbalanced data.
- Almost 17% of the employees didn't continue with the company and left.
- 1991 employees left.
- Almost 83% of the employees continue with the company and didn't leave.
- 10000 employees didn't leave.

imbalanced

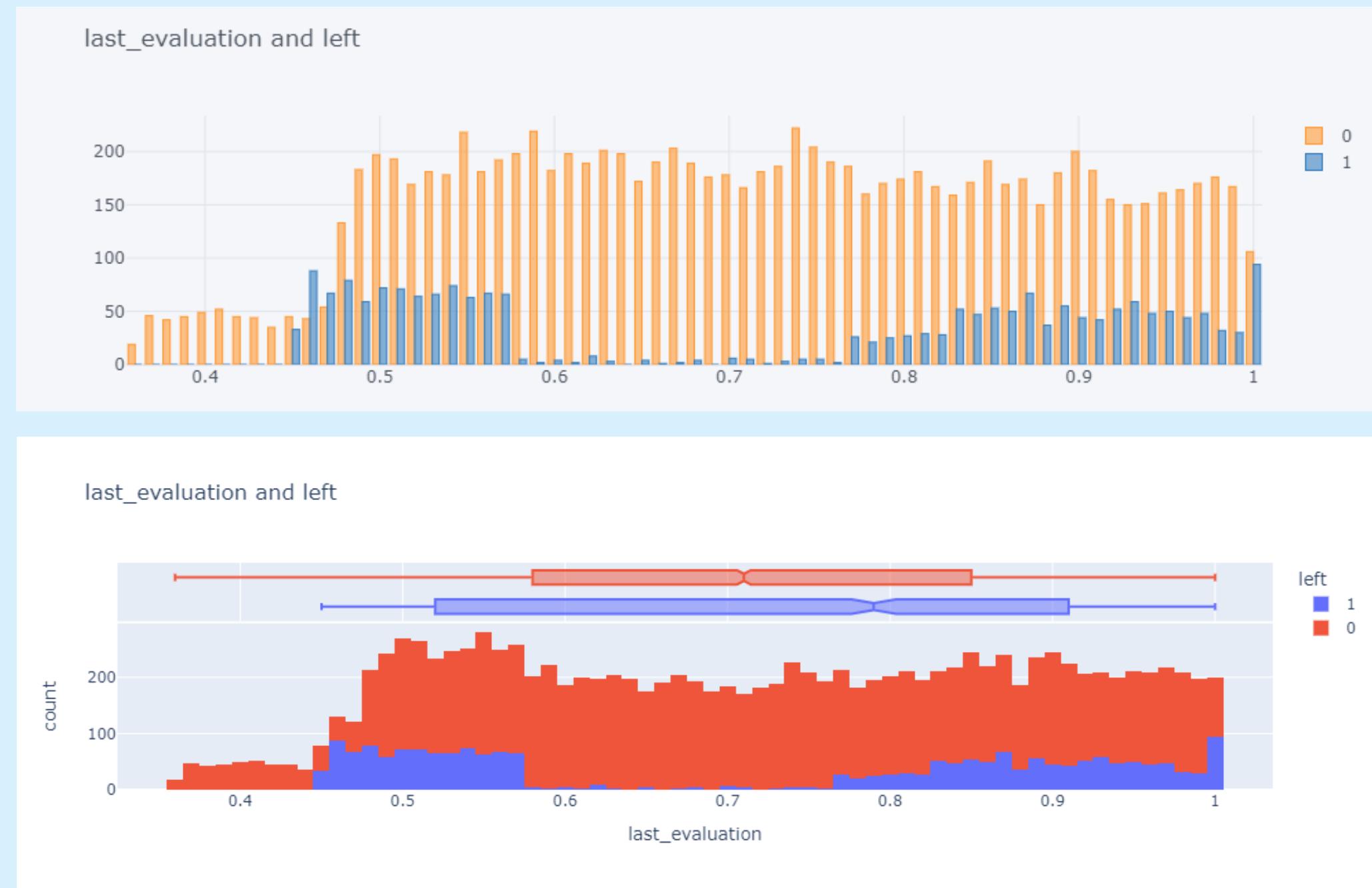


# EDA (satisfaction\_level)



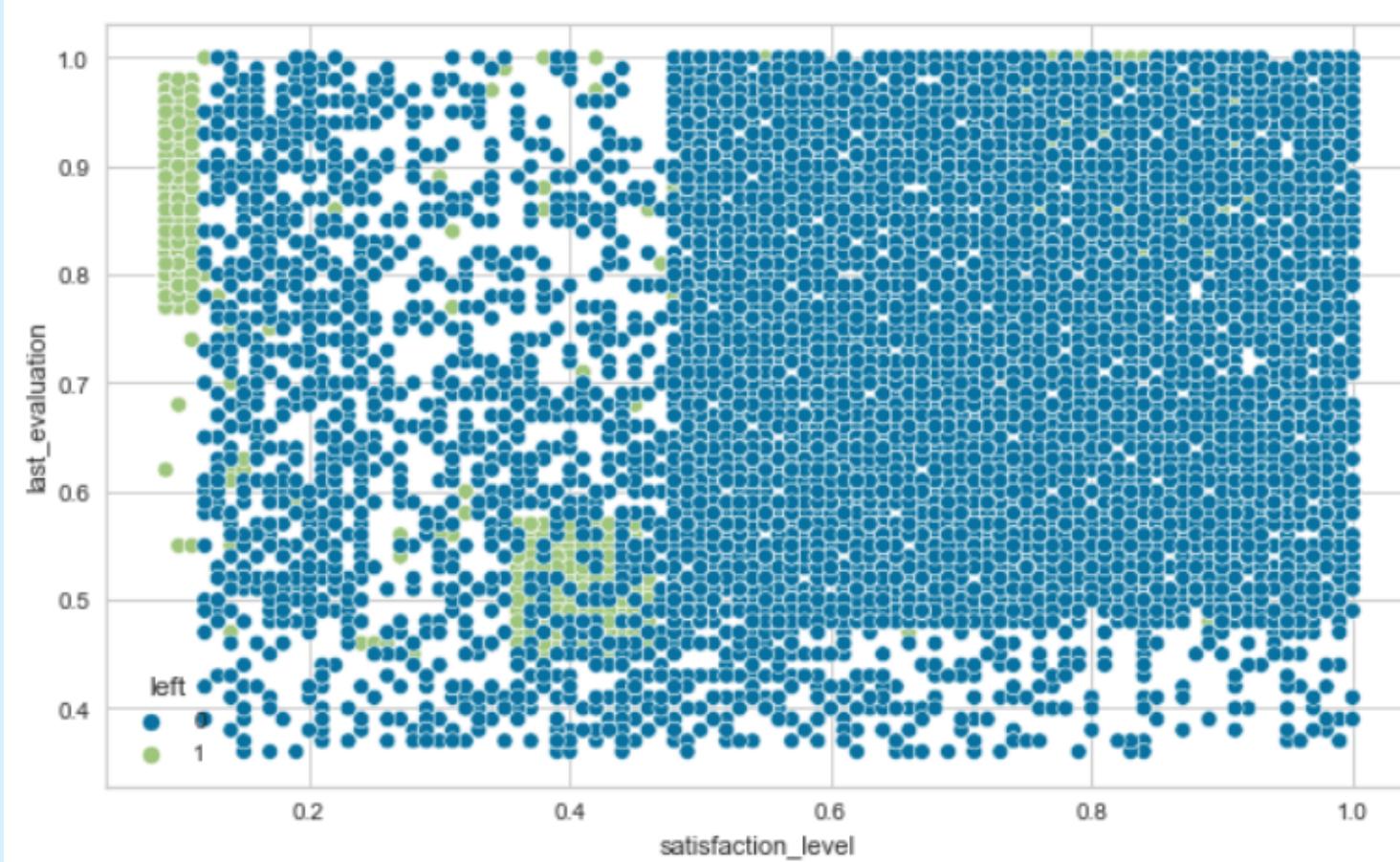
- Normally we expect a low satisfaction level for the employees who have left, so the part near 0 on the x-axis is make sense.
- Besides a group of employees who is not very decisive about their satisfaction level have also been left the company. This group may need extra motivation for employee loyalty. Because they are not so clear in their assessments about their future in the company.
- Also, a group of employees whose satisfaction level is above the average has been left the company. This does not make sense so this must be investigated deeply.
- There may be some other issues:
  - a. The method of gathering this information may be wrong. So the assessment of satisfaction level and the resigning may not be directly proportional.
  - b. The assessment may not be up to date. By the time the satisfaction level may be decreased so at the real-time the satisfaction level of all resigning employees may be close to 0.
  - c. Some of the employees may have hidden their true feelings.

# EDA (last\_evaluation)



- Most of the employees have been assessed above 0.45.
- The evaluation of the resigning employees is gathered in two groups. The first is around 0.5 and the second is between 0.8 and 1.0.
- Intensive work may cause the resign of high evaluated employees (second group). Because the employer will be happy with the performance of these staff, however, it will be a burden for employees.

# EDA (last\_evaluation)



It becomes meaningful when the satisfaction level of employees and the evaluation of the employer brings together.

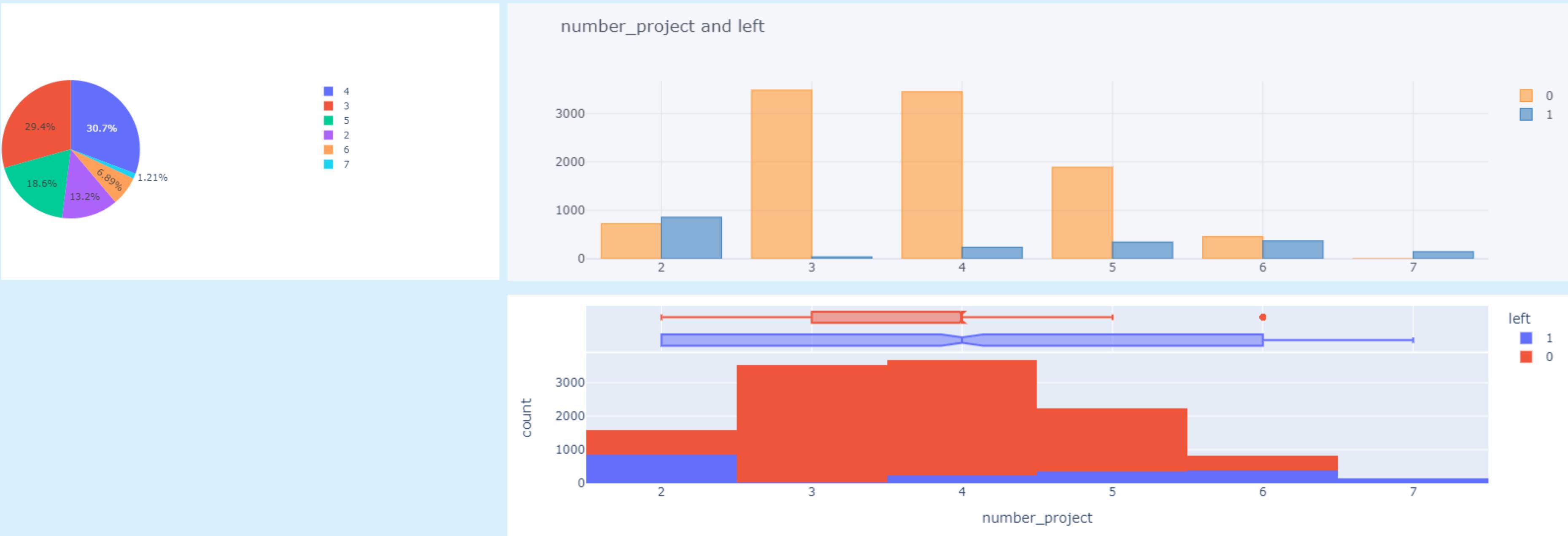
As seen in the graph; the resigning employees are grouped into two different clusters.

1. The first group has a satisfaction level of 0.4 and the last evaluation of 0.5. This group does not have a clear idea about the company and the employer does not have a clear assessment as well. Other features affecting this group have to be investigated. What are the main problems of this group? Why are they confusing? What are the pros and cons of the company for these groups? and so on...

2. The second group has low satisfaction even if the employer evaluated them with high degrees. Then what can be the main problem of this group?

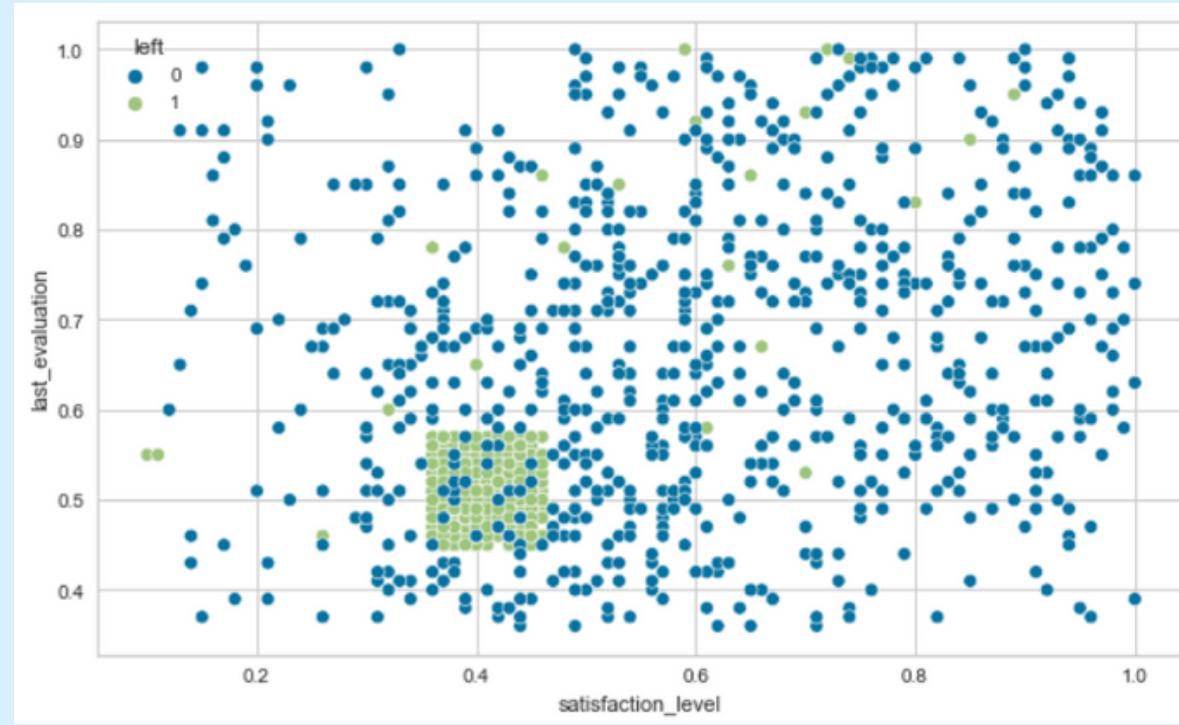
***Intensive work with a low salary may affect this group. Or intensive work without promotion may cause. In the next steps, the workload and motivation factors of this group have to be investigated.***

# EDA (number\_project)

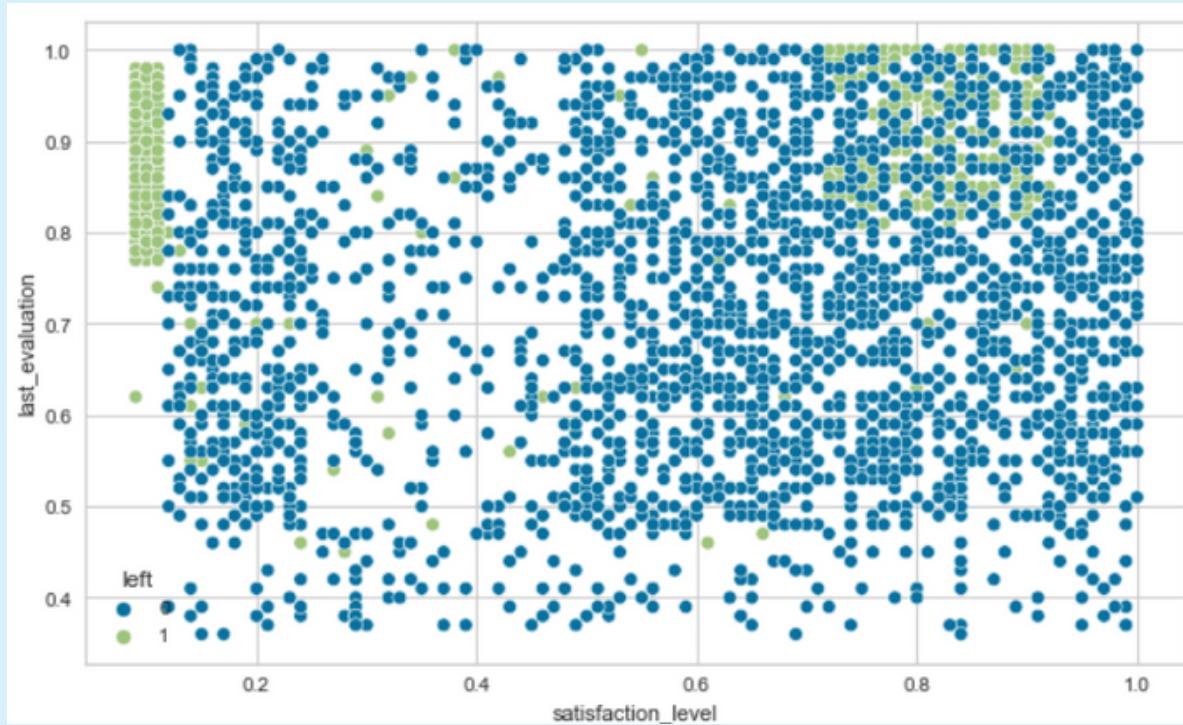


- The number of leaving employees is higher among those who have only two projects during the period. This can be summed up as: "**the employees with only two projects feel worthless or emptied**". Because most of the employees work on three or four projects.
- With the 6th project, the number of resignings is getting closer to the number of ongoing staff. There are no ongoing staff members who were assigned to 7 projects.
- Working on more projects may cause an intensive workload, regarding to this the satisfaction level may decrease with the insufficient motivators.

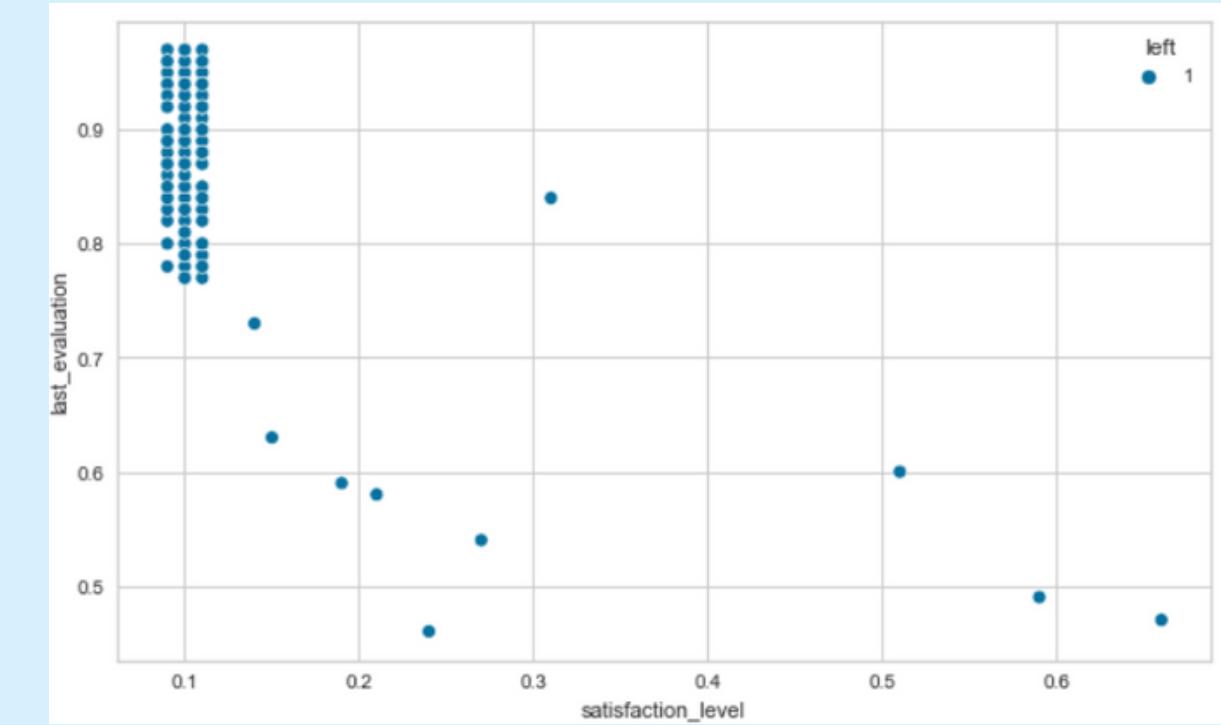
# EDA (number\_project)



number\_project = 2



number\_project > 4

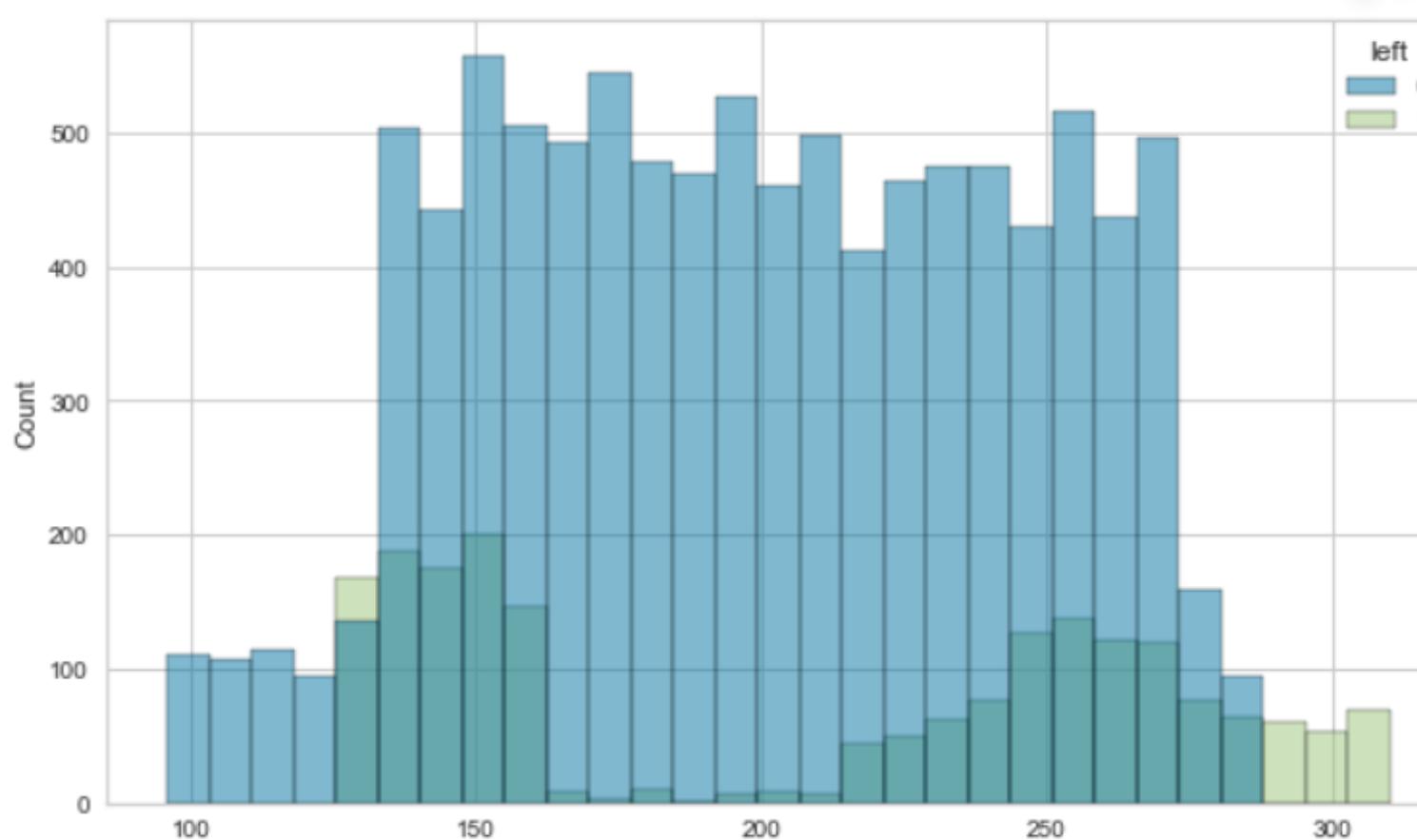


number\_project = 7

If we look at the satisfaction level, evaluation score, and the number of projects together;

1. The group of undecideds who were evaluated as 0.5 is the group who worked on the only ***two projects***. As a result, our hypothesis about this group is becoming more clear. As the employer does not assign enough projects to this group, he/she cannot evaluate their performance and they feel worthless. Therefore, they are unsure about their future in the company. This may lead them to leave.
2. The leaving employees who worked on more than four projects are the ***second group of the last\_evaluation section***.
3. The second group of last\_evaluation section are mostly worked on seven projects and left the company. So again our hypothesis about this group is now more definite.

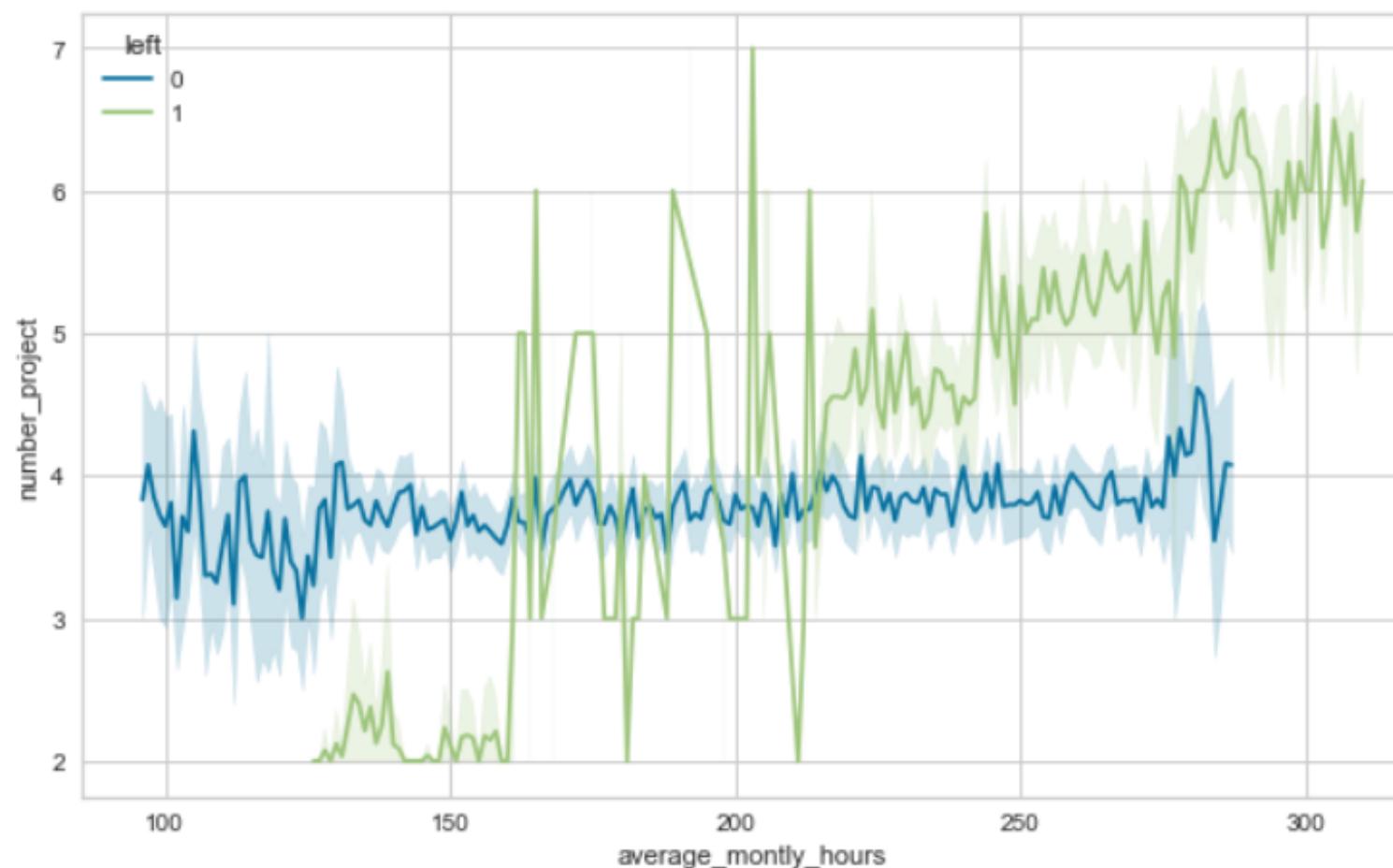
# EDA (average\_monthly\_hours)



According to average monthly working hours again we have two peaks. One is around 140 hours and the other is around 260 hours.

More than 280 hours there is no ongoing staff.

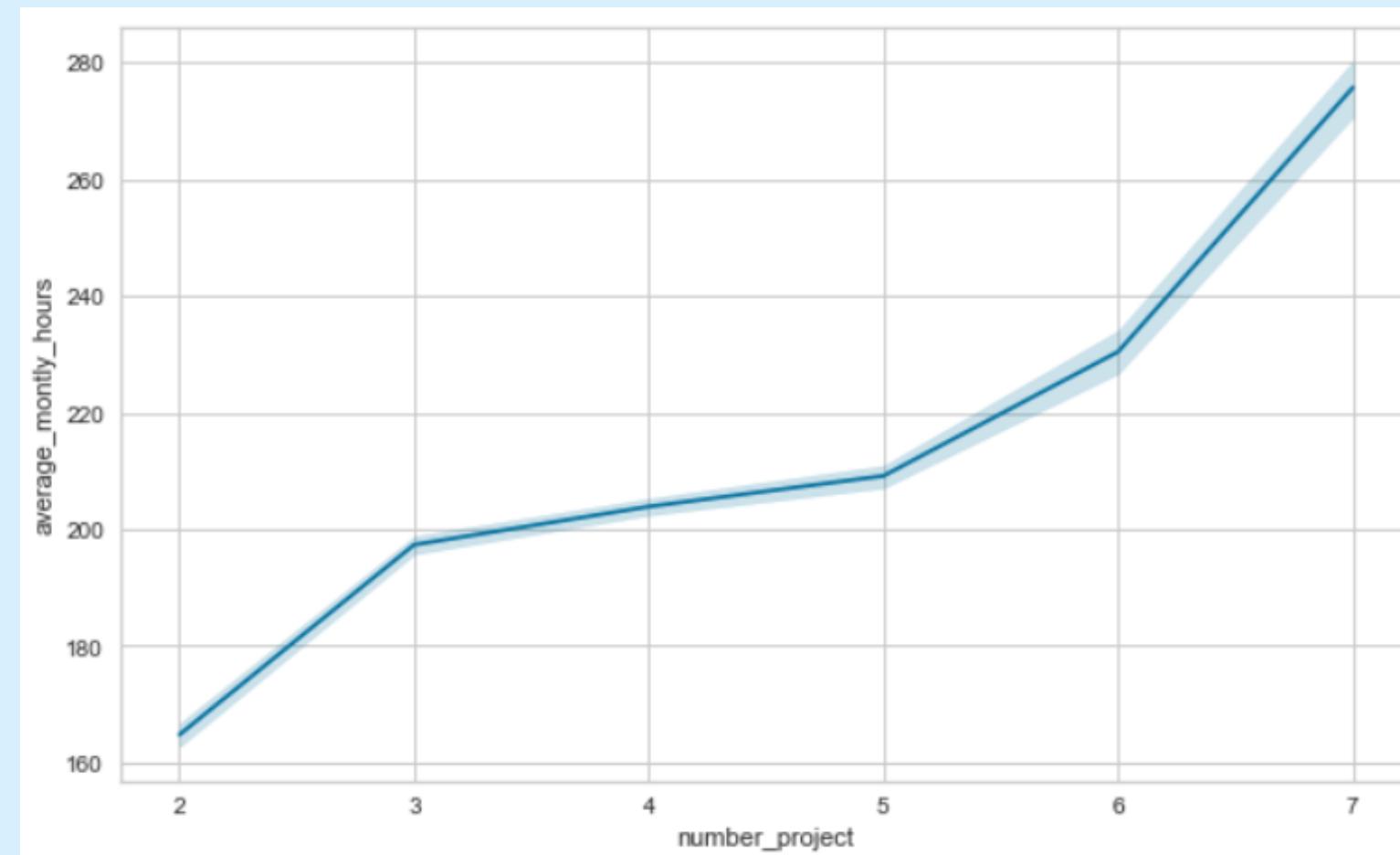
So the next question is "The average monthly working hours are related to project number or not?"



In the graph, it is seen that the group working on two projects is working nearly 130-160 hours monthly. It can be assessed that they have only two simple projects that they don't need to work hard, so their loyalty is weak.

Most of the employees are working 135-275 hours monthly. In this group usually, the employees who get two or more than five projects leave the company. With the increasing number of projects, the average monthly working hours and the number of resigning are increasing.

# EDA (average\_monthly\_hours)



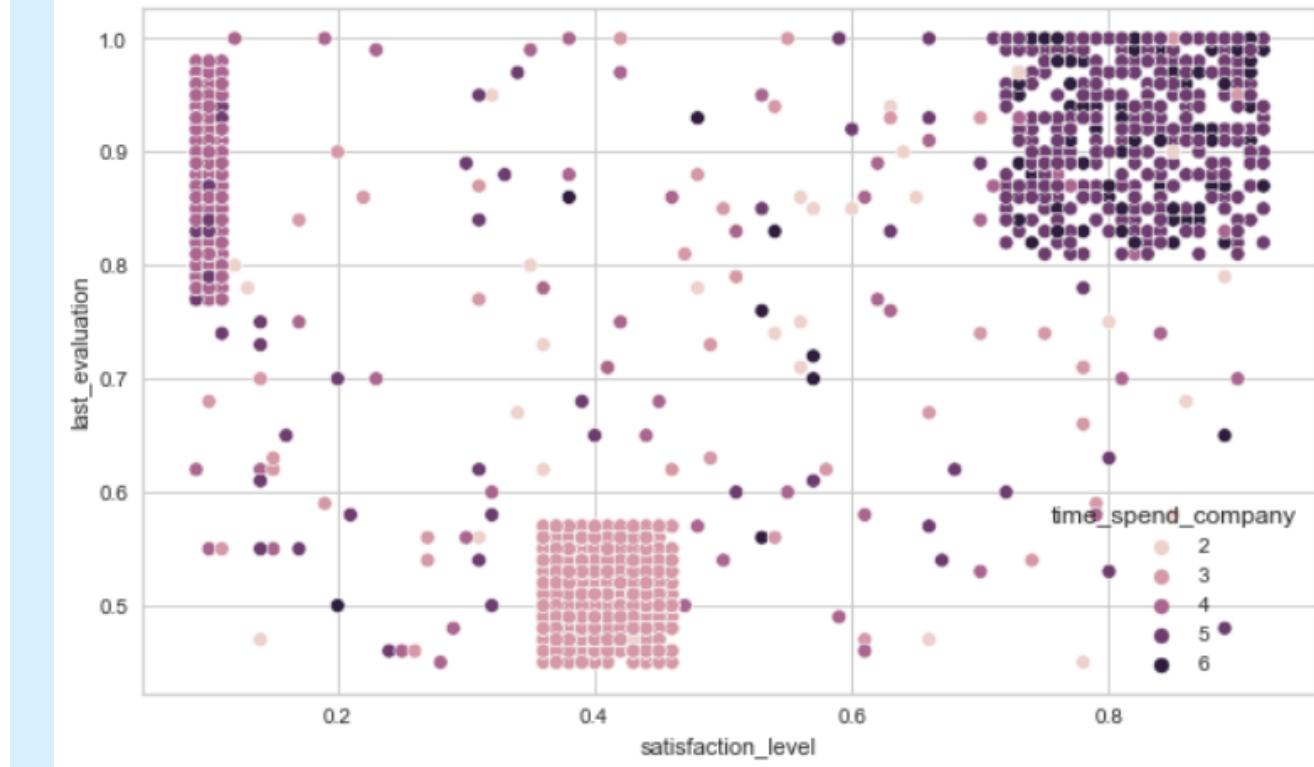
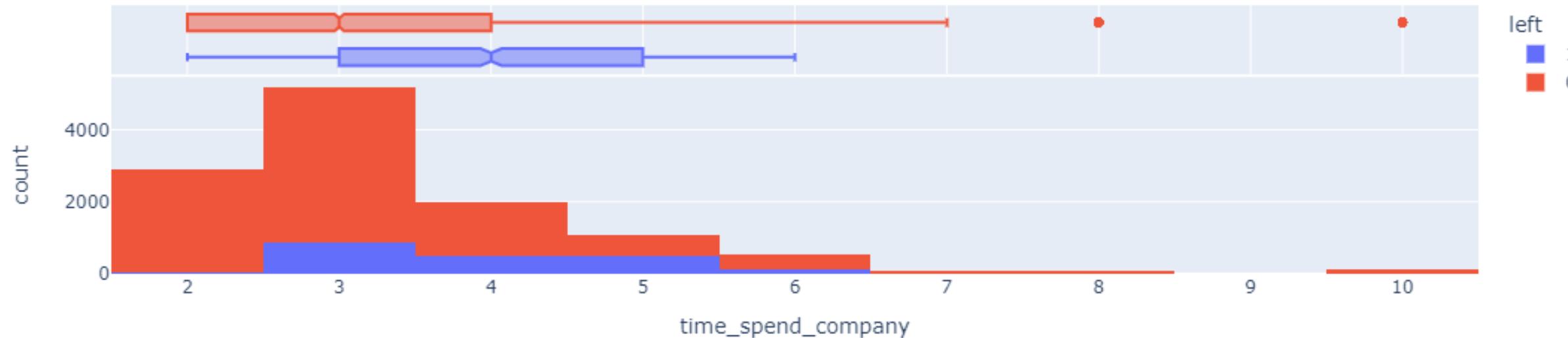
The increase of the average monthly hours according to the number of projects is seen on the graph.

The rate of increase is higher between two and three projects, and after five projects. So it clearly defines the number of resigning due to the working hours.

***There needs to be an adjustment about the project numbers, working hours, and workload. The projects must be assigned to more employees. Also, better incentives must be offered to staff who are working hard.***

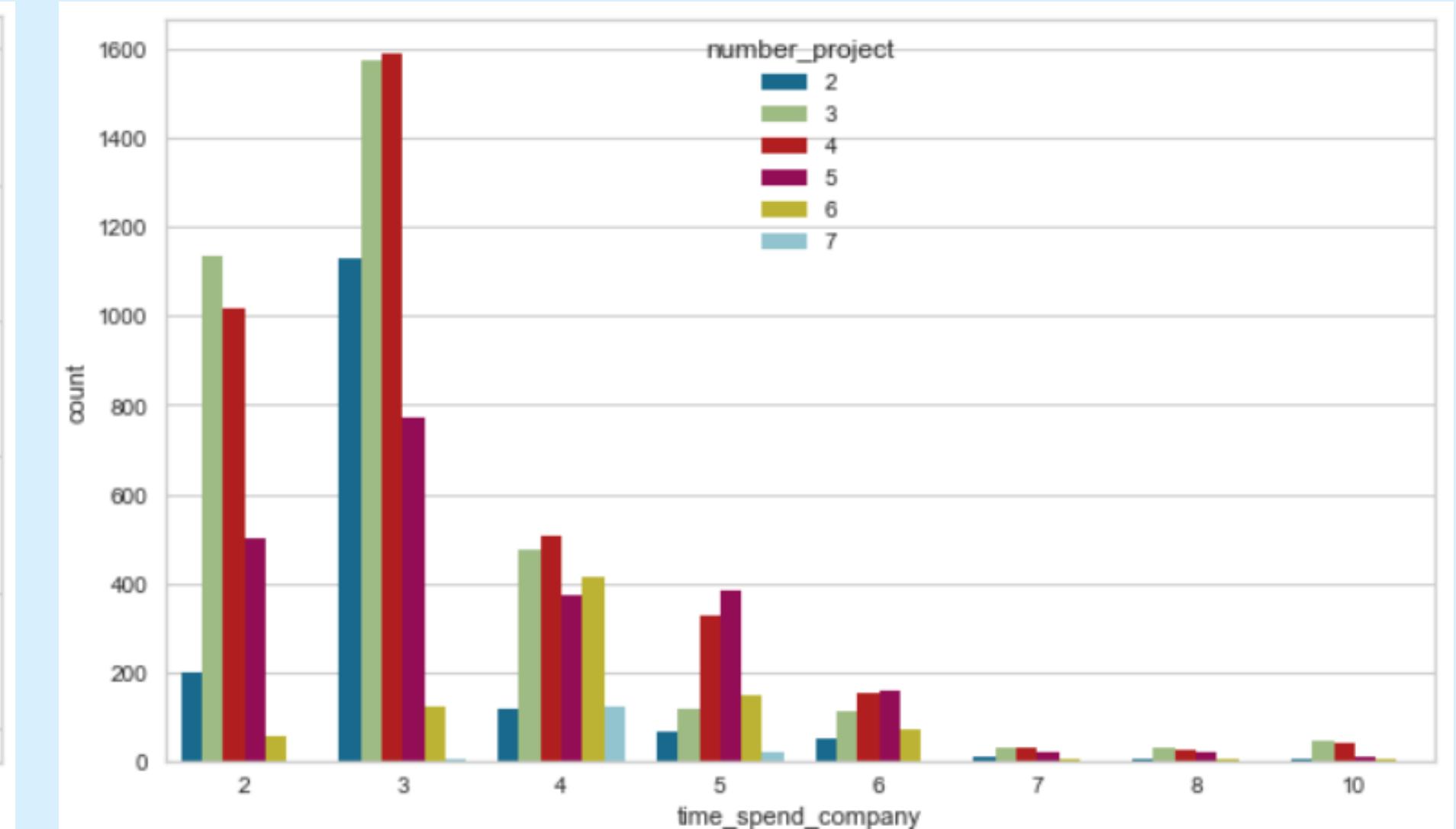
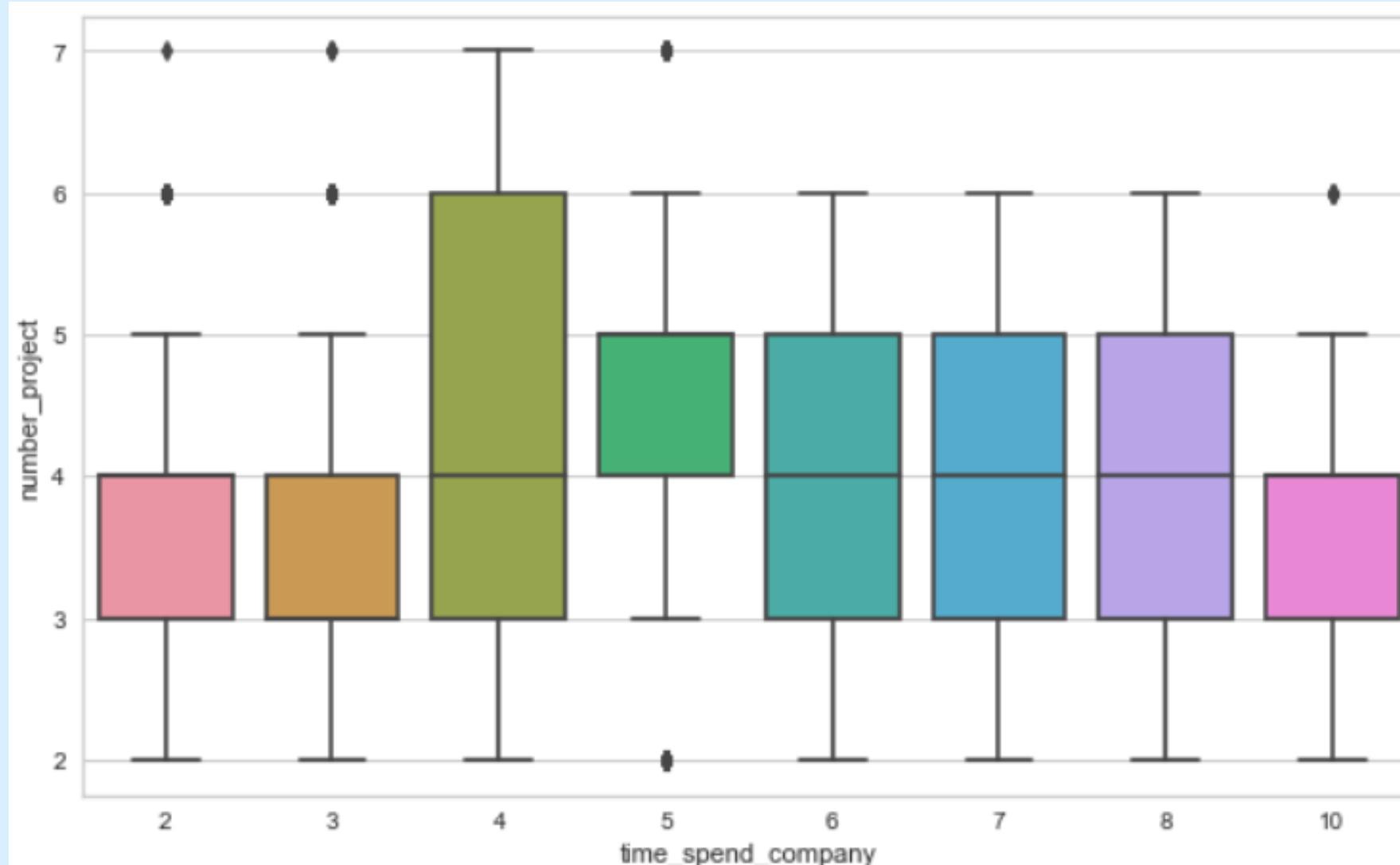
# EDA (time\_spend\_company)

time\_spend\_company and left



- Most of the resigning employees are in their third year. By the time their intention to resign is being decreased.
- As can be seen on the graph, the employees are not able to make a clear assessment of the company during the first three years of their employment. This, coupled with the other factors, tends to lead to leaving the company after three years.
- By the fourth year, their workload increases and their satisfaction decreases.
- After the fifth year, they make an assessment, "they will leave or not".
- If they decide to continue in the company, they never consider leaving after the sixth year.

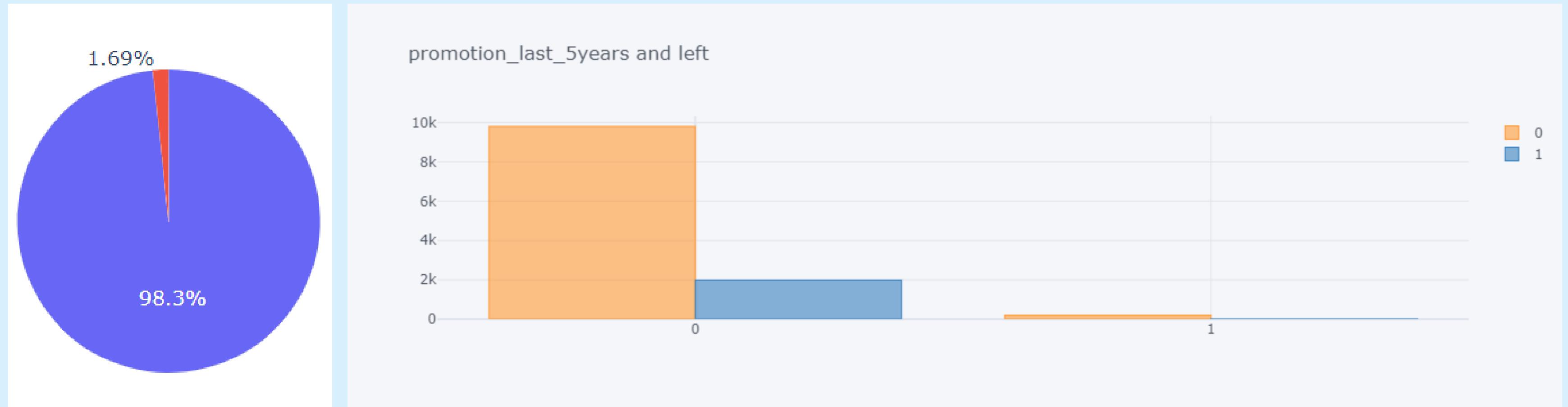
# EDA (time\_spend\_company)



**Then how is the relationship between workload and time spent in the company?**

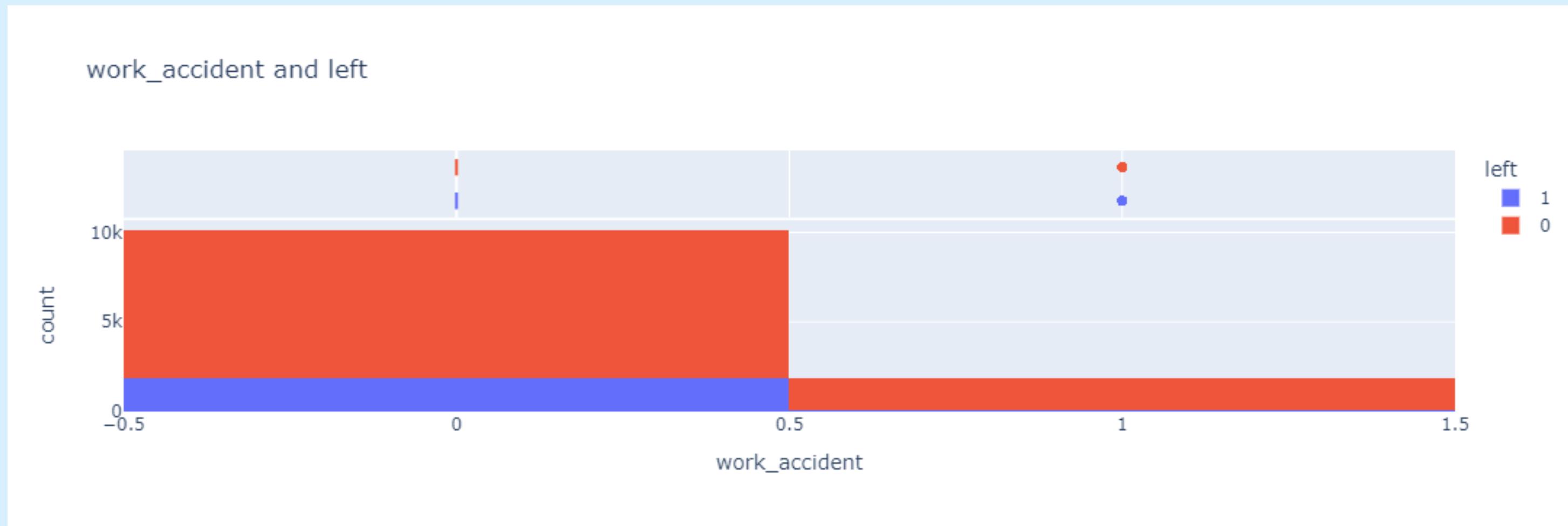
The third-year staff has the most workload. After that year number of participating projects is decreasing stepped. It makes sense. The experienced staff becoming team leader or manager position. That's why fewer of them can be assigned to projects.

# EDA (promotion\_last\_5years)



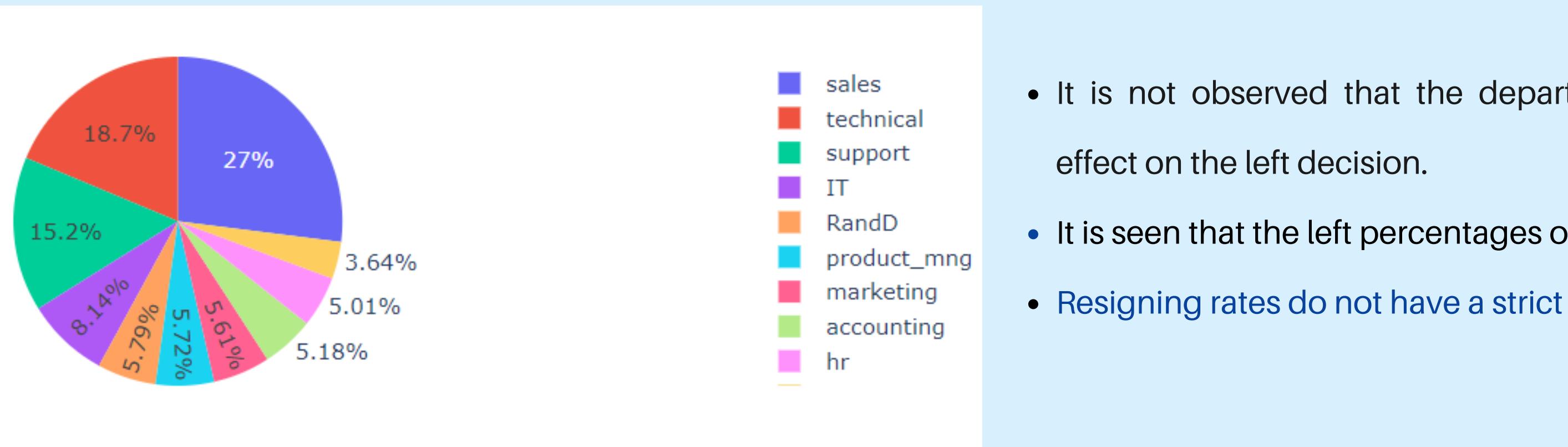
- Receiving a promotion in the last 5 working years is not determinative in terms of leaving or continuing to work.
- However, the percentage of those who receive promotions, even if it is small, is higher than those who do not.
- **Resigning rates do not have a strict relation with getting promoted.**

# EDA (work\_accident)

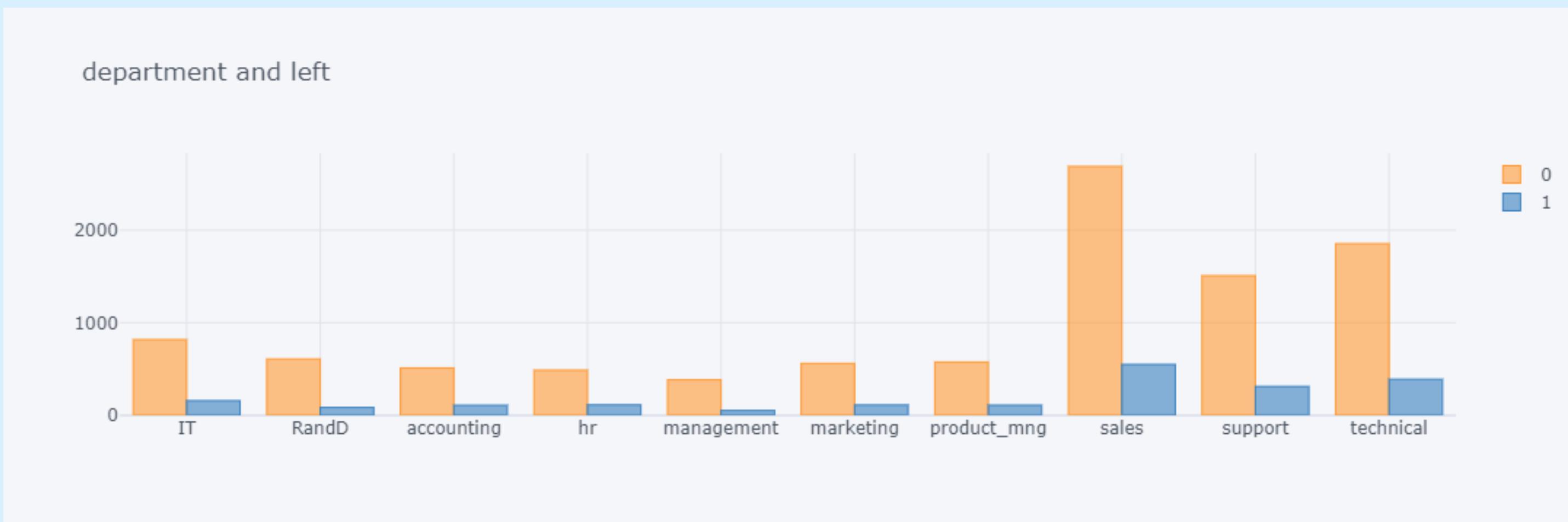


- Left ratios are similar between those who have had a work accident and those who have not.
- It does not appear to be a determining factor. In fact, it can be said that the left rate of those who have had a work accident is proportionally lower.
- Resigning rates do not have a strict relation with work accidents.

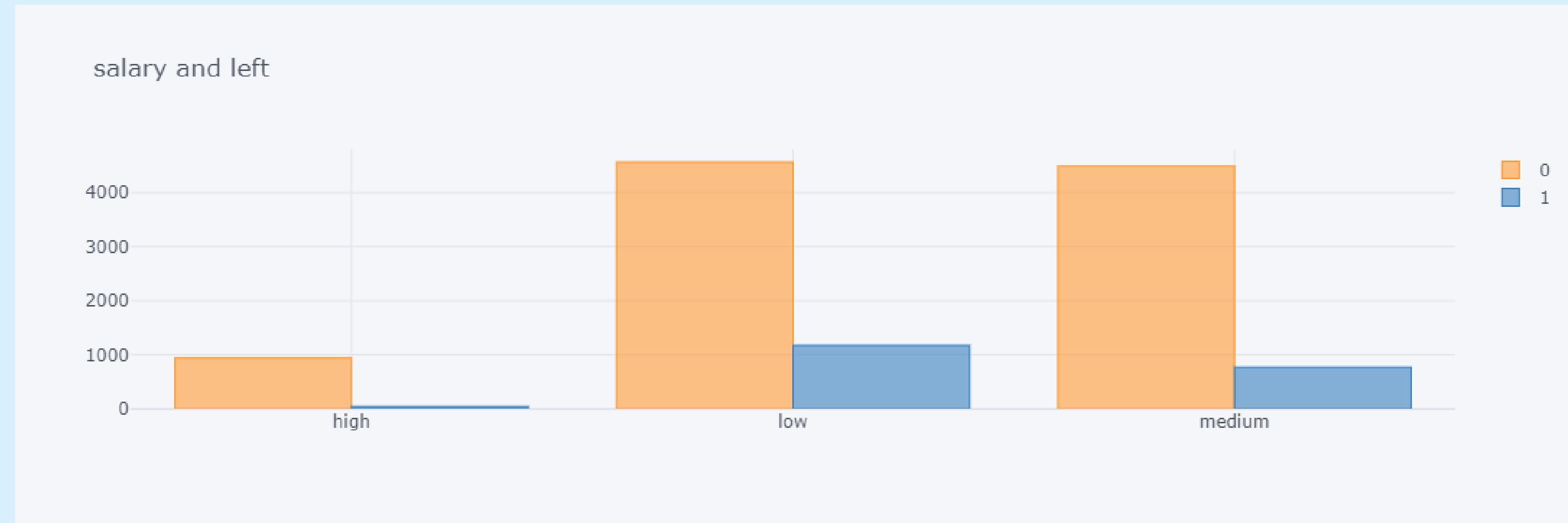
# EDA (Departments )



- It is not observed that the departments worked alone have an effect on the left decision.
- It is seen that the left percentages of the departments are similar.
- Resigning rates do not have a strict relation with work accidents.

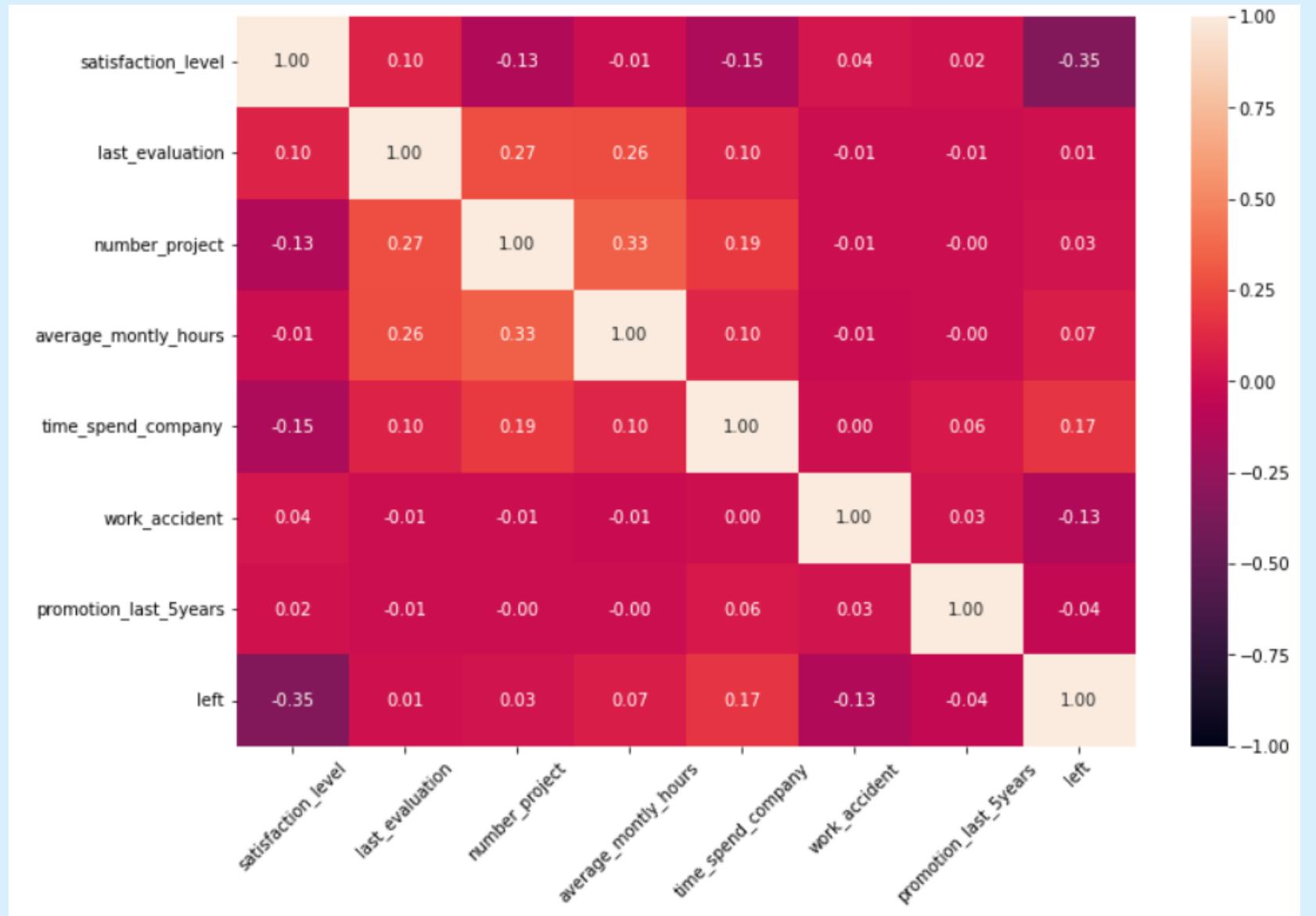


# EDA (salary)

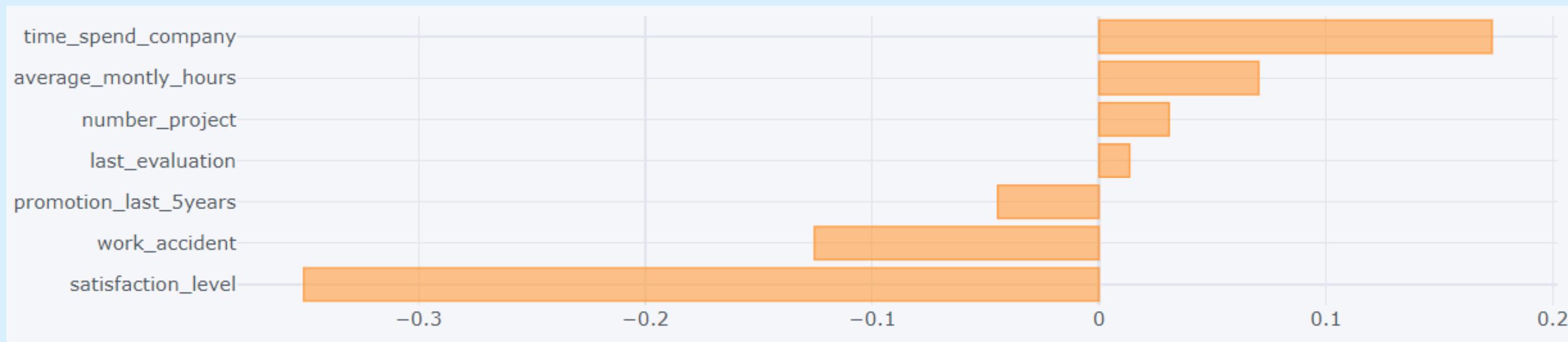


- Nearly 1/3 of the employees who have a salary level of low has left.
- It is seen that the left percentages of the salary are similar.
- Even if it is small, there is an increase in the form of high-medium-low according to the salary status.

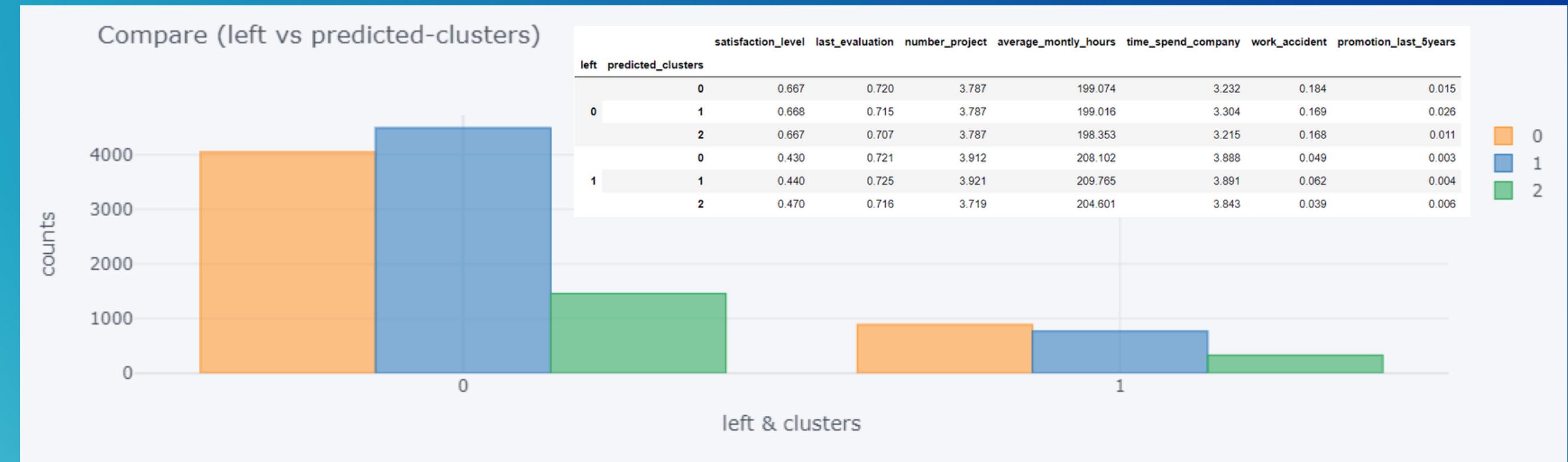
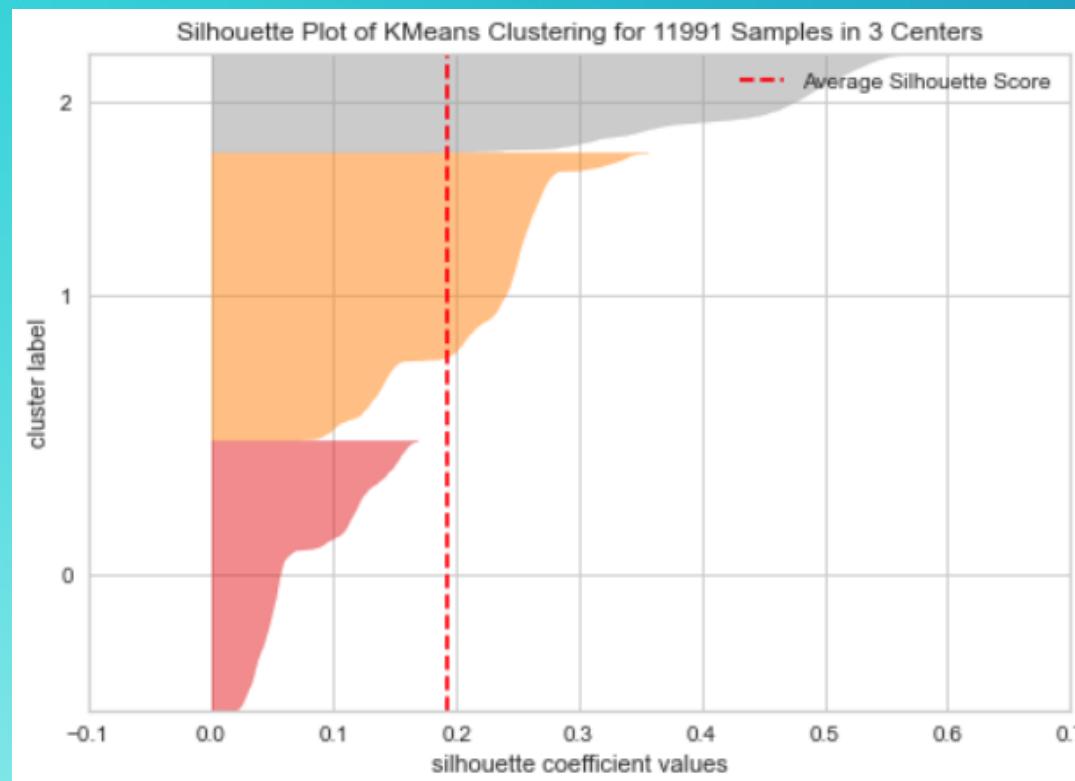
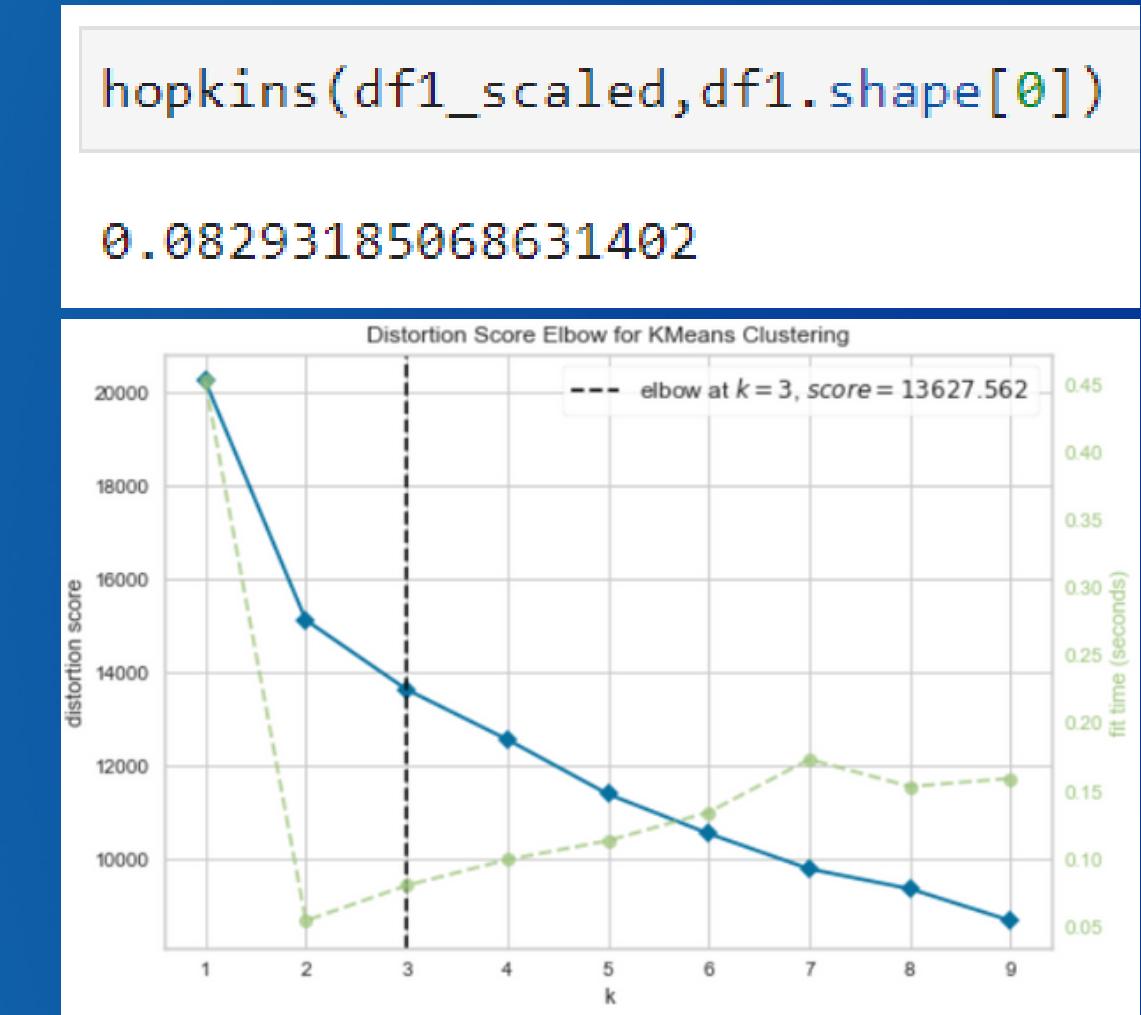
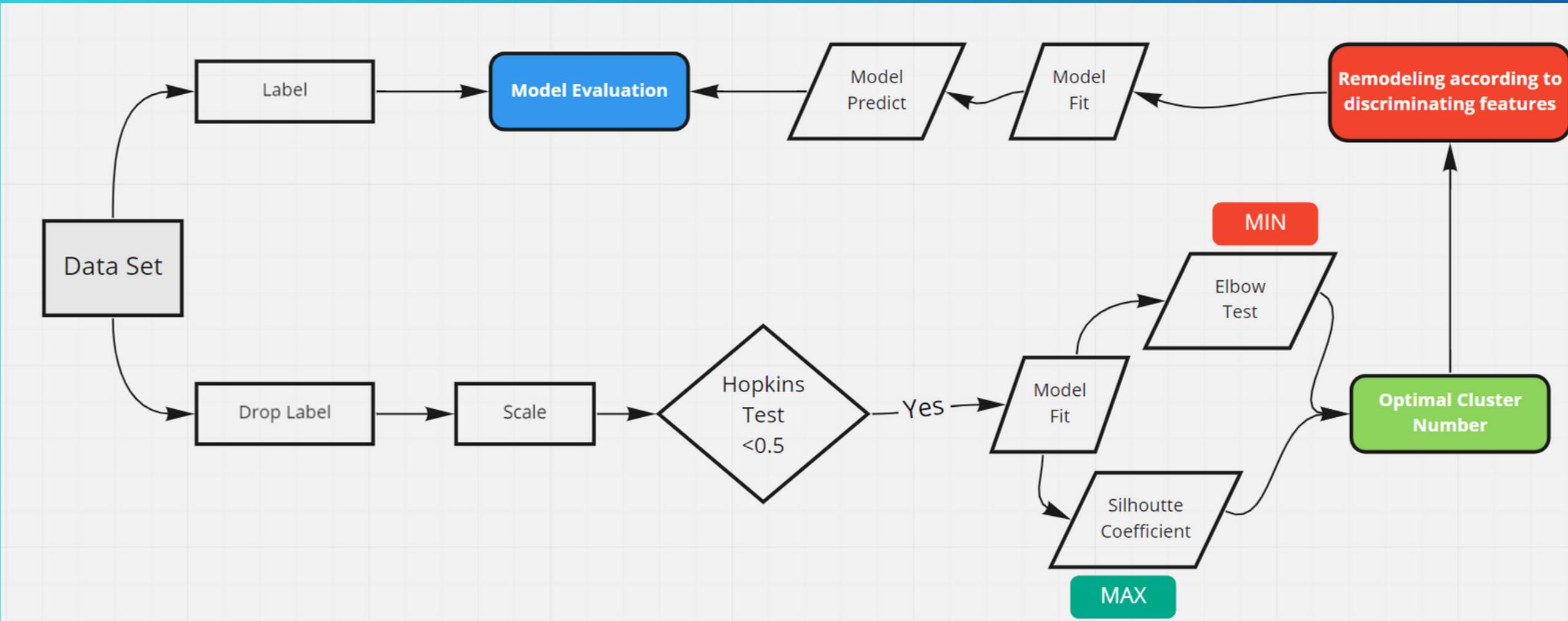
# EDA (CONCLUSION)



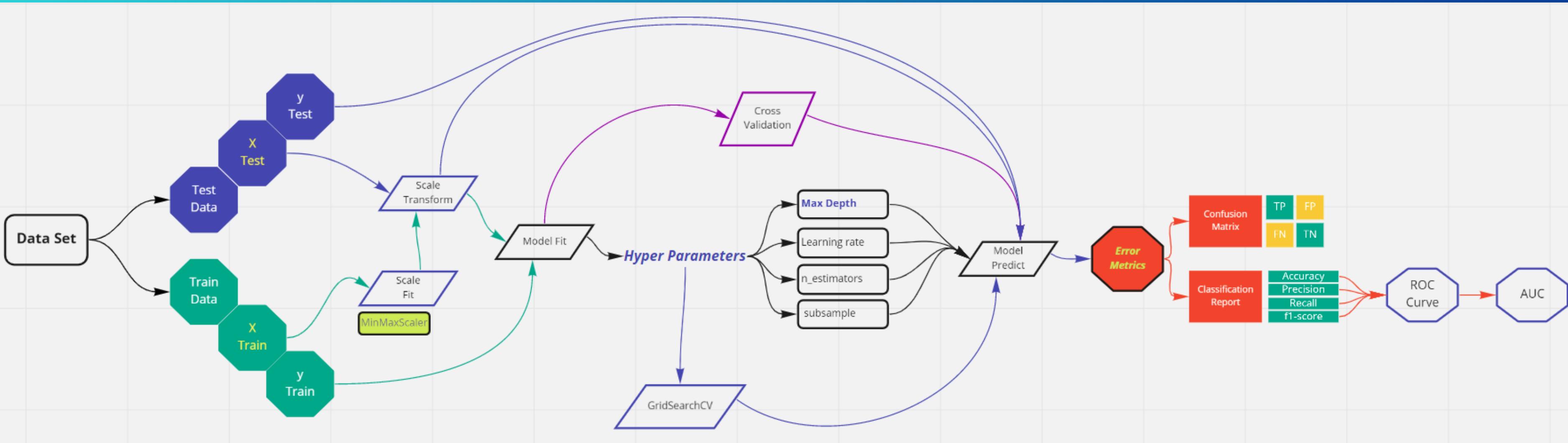
- There is no multicollinearity problem among the features.
- We have weak level correlation between the numerical features and the target column.
- Also there is weak level correlation between the columns.



# K-MEANS



# GRADIENT BOOSTING



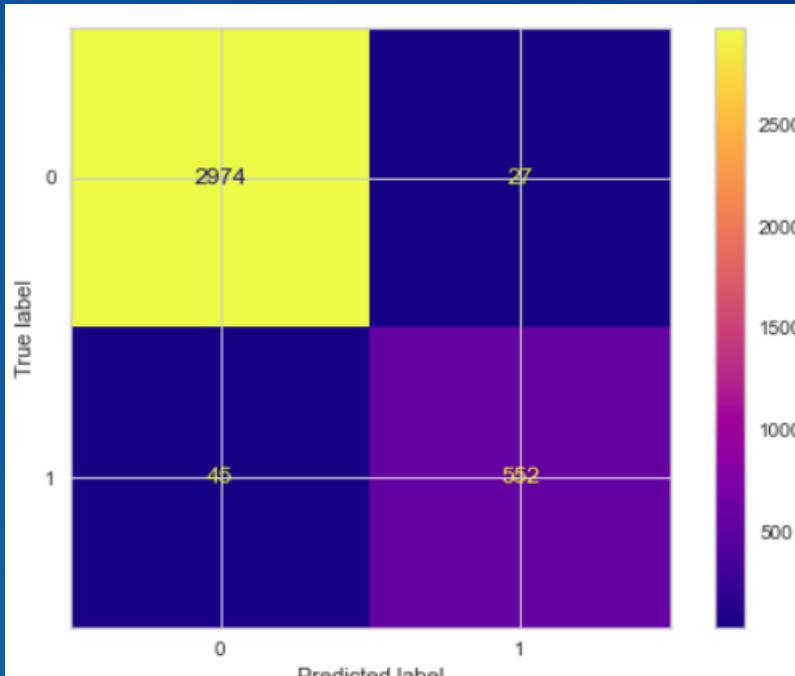
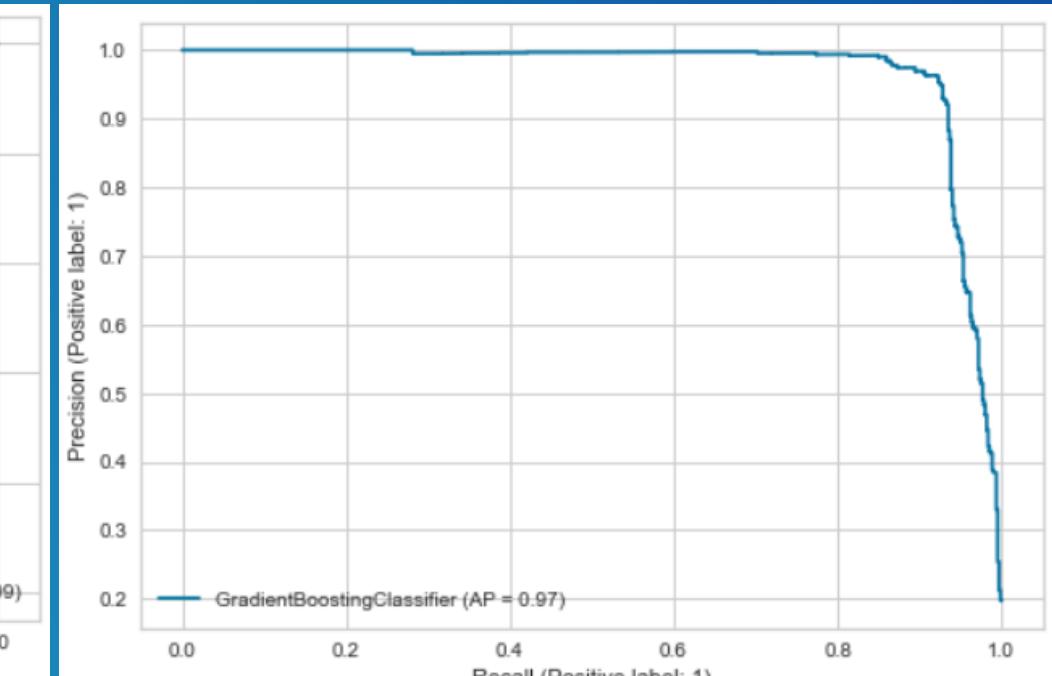
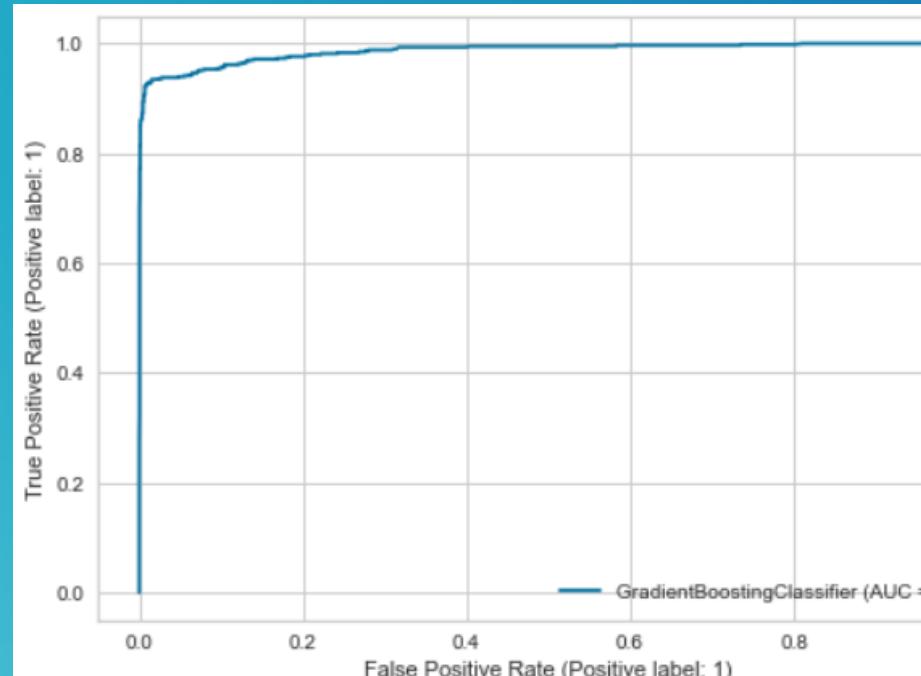
	train_set	test_set
Accuracy	0.983	0.980
Precision	0.969	0.953
Recall	0.928	0.925
f1	0.948	0.939

	train_set	test_set
Accuracy	0.986	0.985
Precision	0.989	0.987
Recall	0.926	0.920
f1	0.957	0.952

GB Model

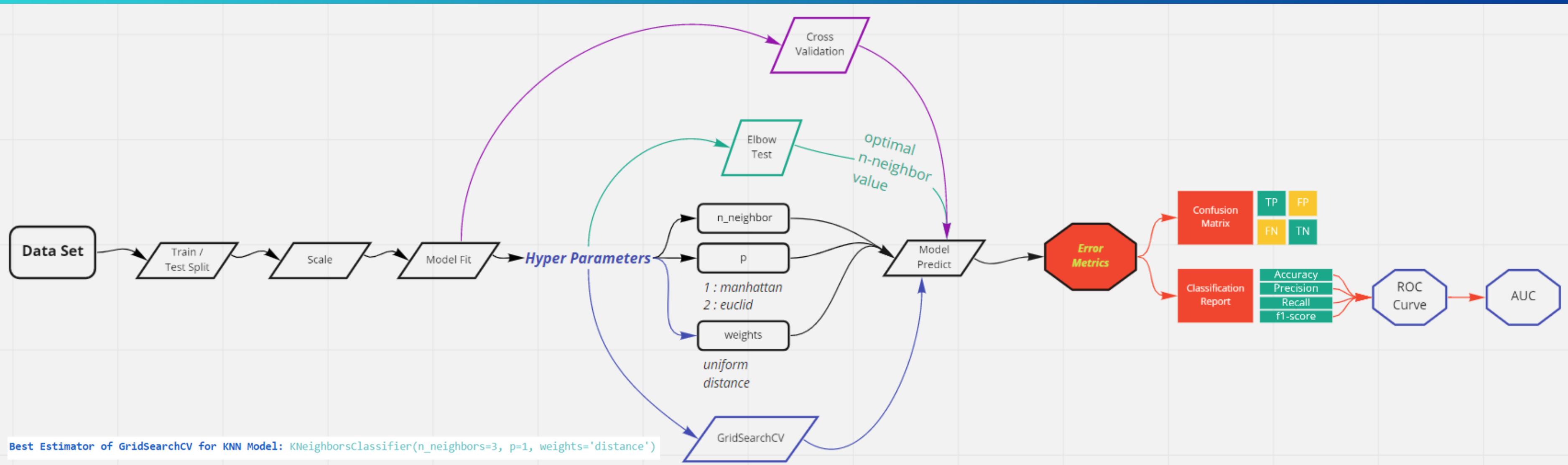
Gridsearch CV

test_accuracy	0.981
test_precision	0.965
test_recall	0.922
test_f1	0.943
test_roc_auc	0.985



CV

# KNN

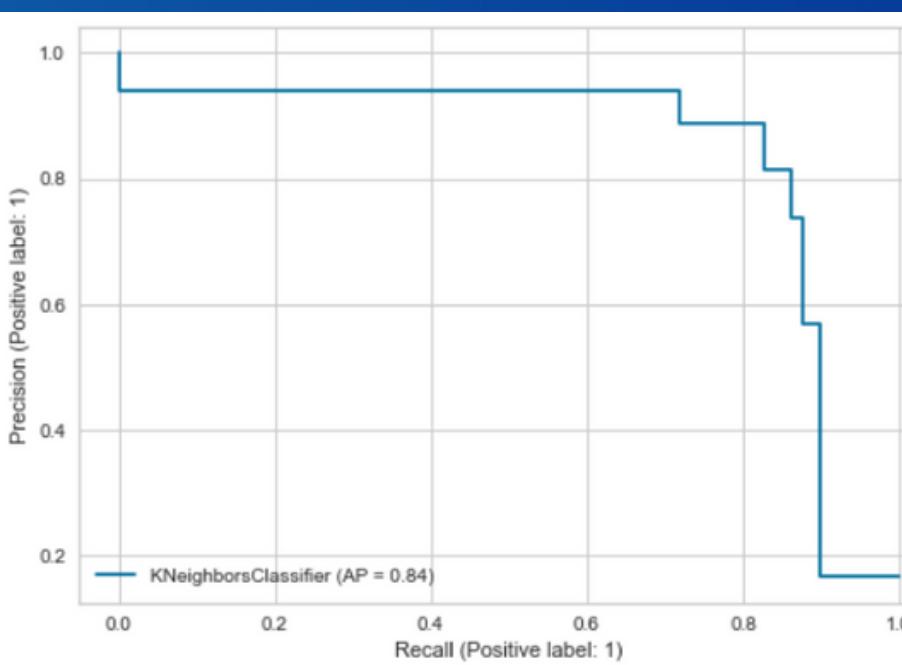
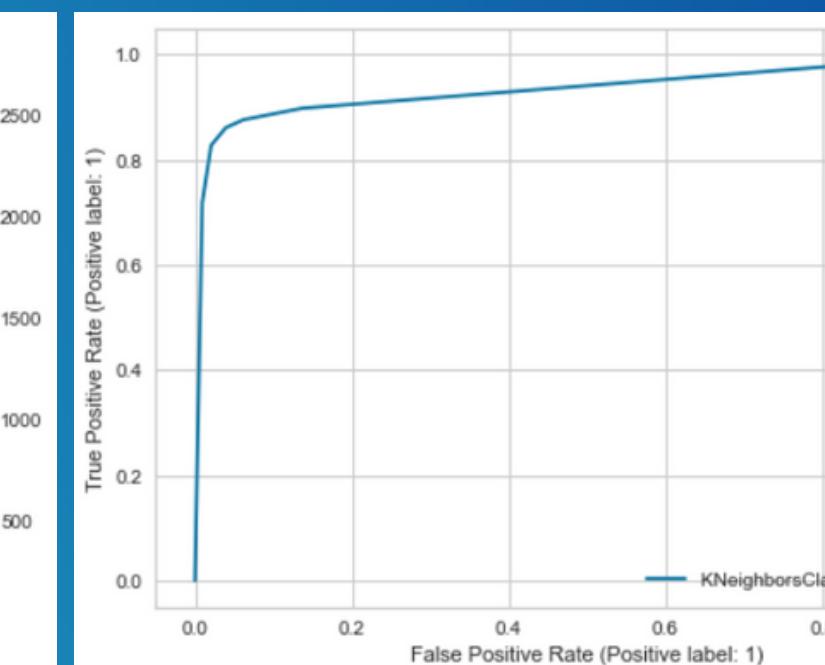
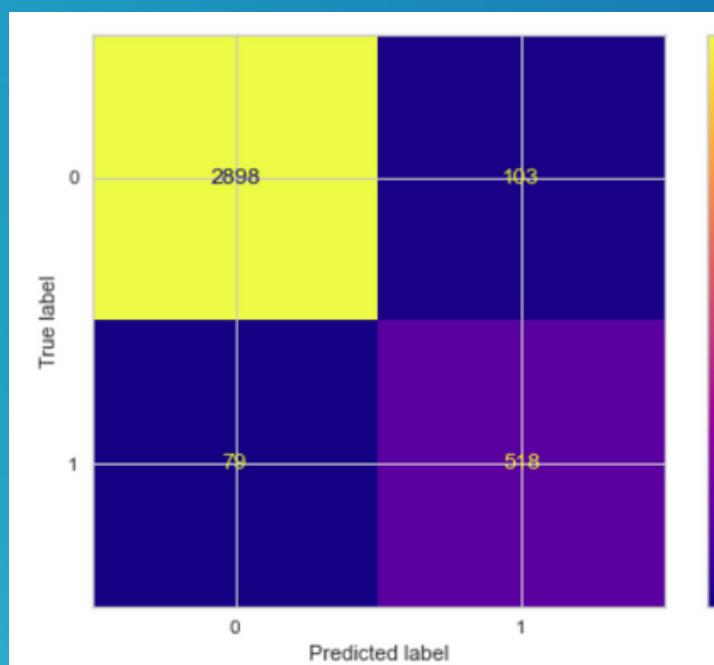
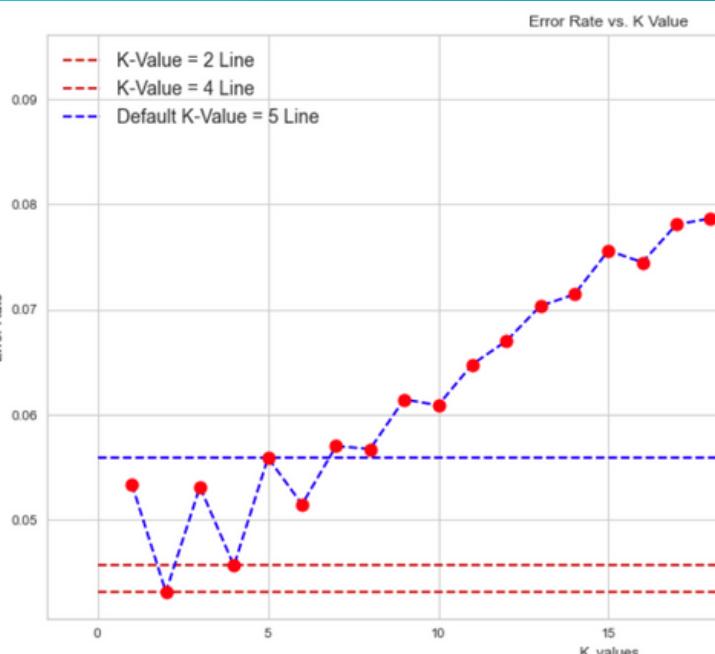


	train_set	test_set
<b>Accuracy</b>	0.960	0.944
<b>Precision</b>	0.879	0.813
<b>Recall</b>	0.878	0.861
<b>f1</b>	0.879	0.836

## KNN Model

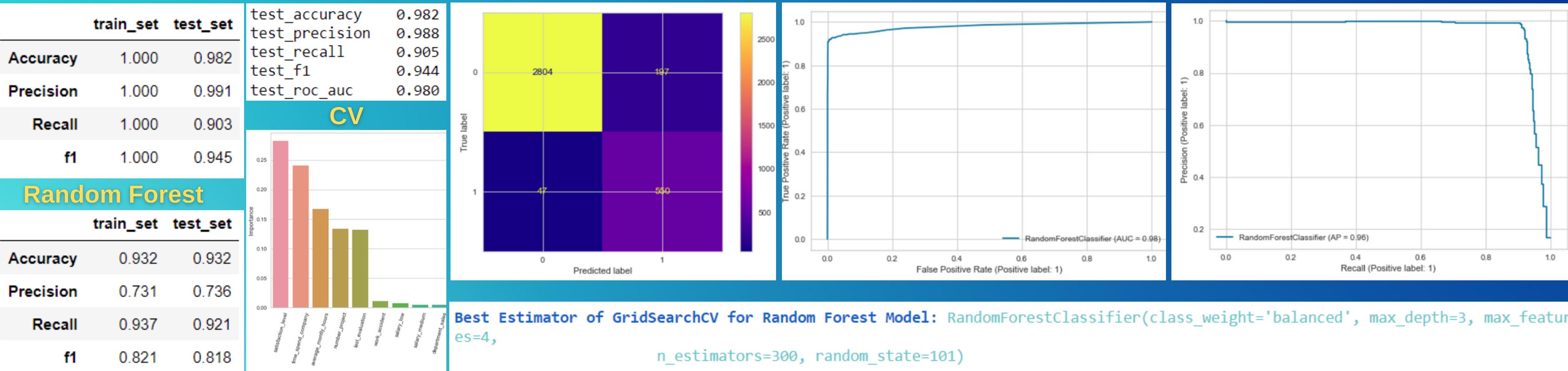
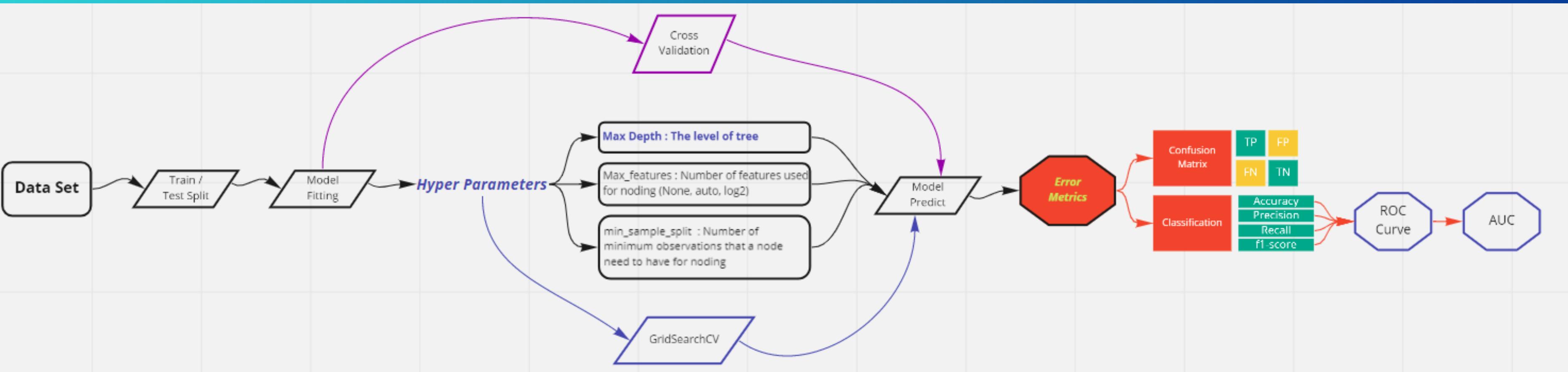
```

test_accuracy      0.948
test_precision     0.835
test_recall        0.859
test_f1            0.846
test_roc_auc       0.943
  
```

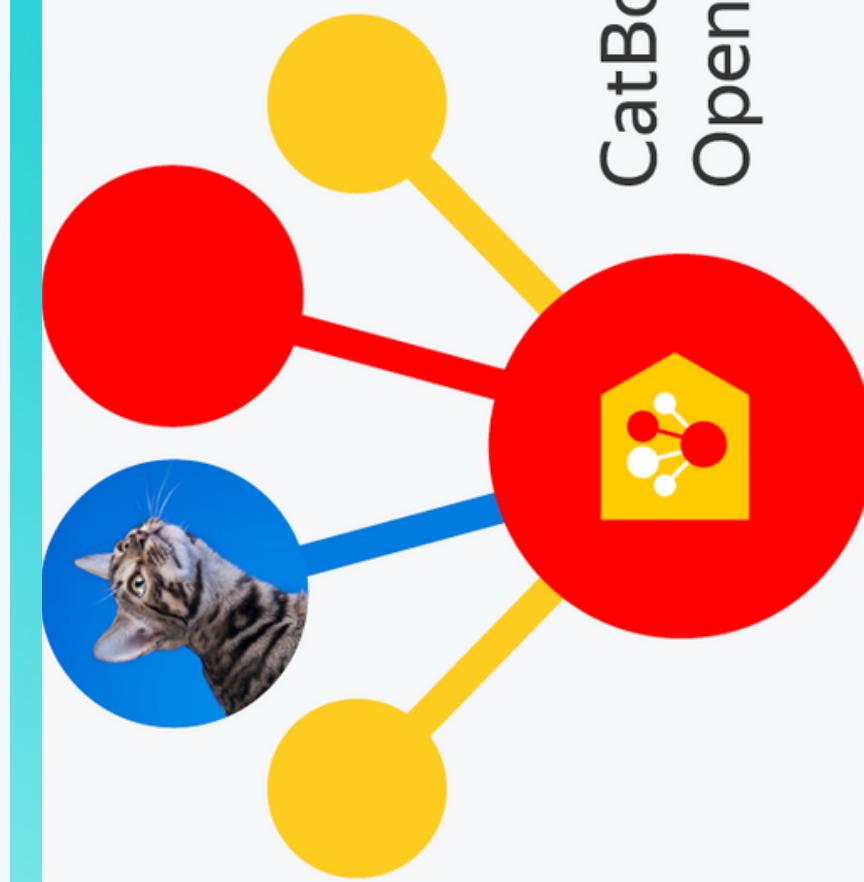


CV

# RANDOM FOREST



# CATBOOST



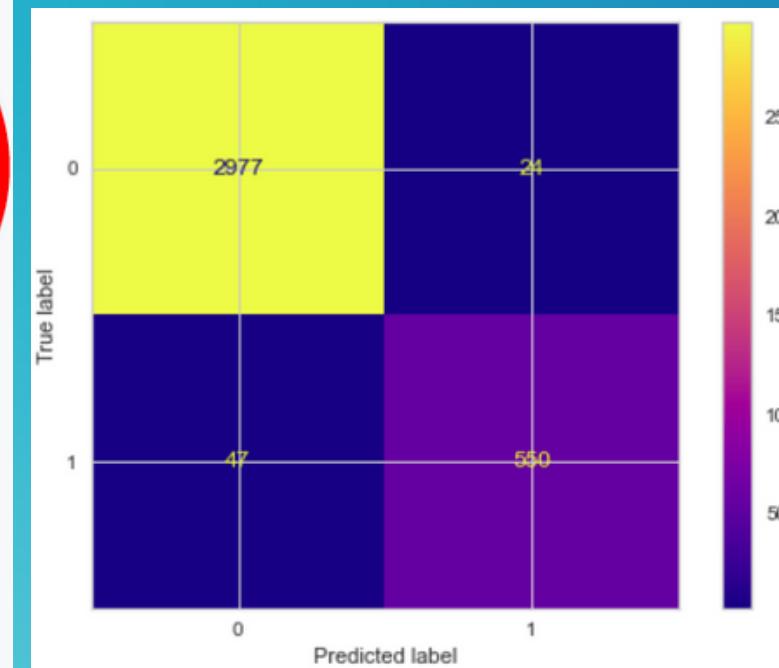
CatBoost  
Open-source ML library

	train_set	test_set
<b>Accuracy</b>	0.994	0.980
<b>Precision</b>	0.977	0.958
<b>Recall</b>	0.987	0.921
<b>f1</b>	0.982	0.939

## Catboost Classifier

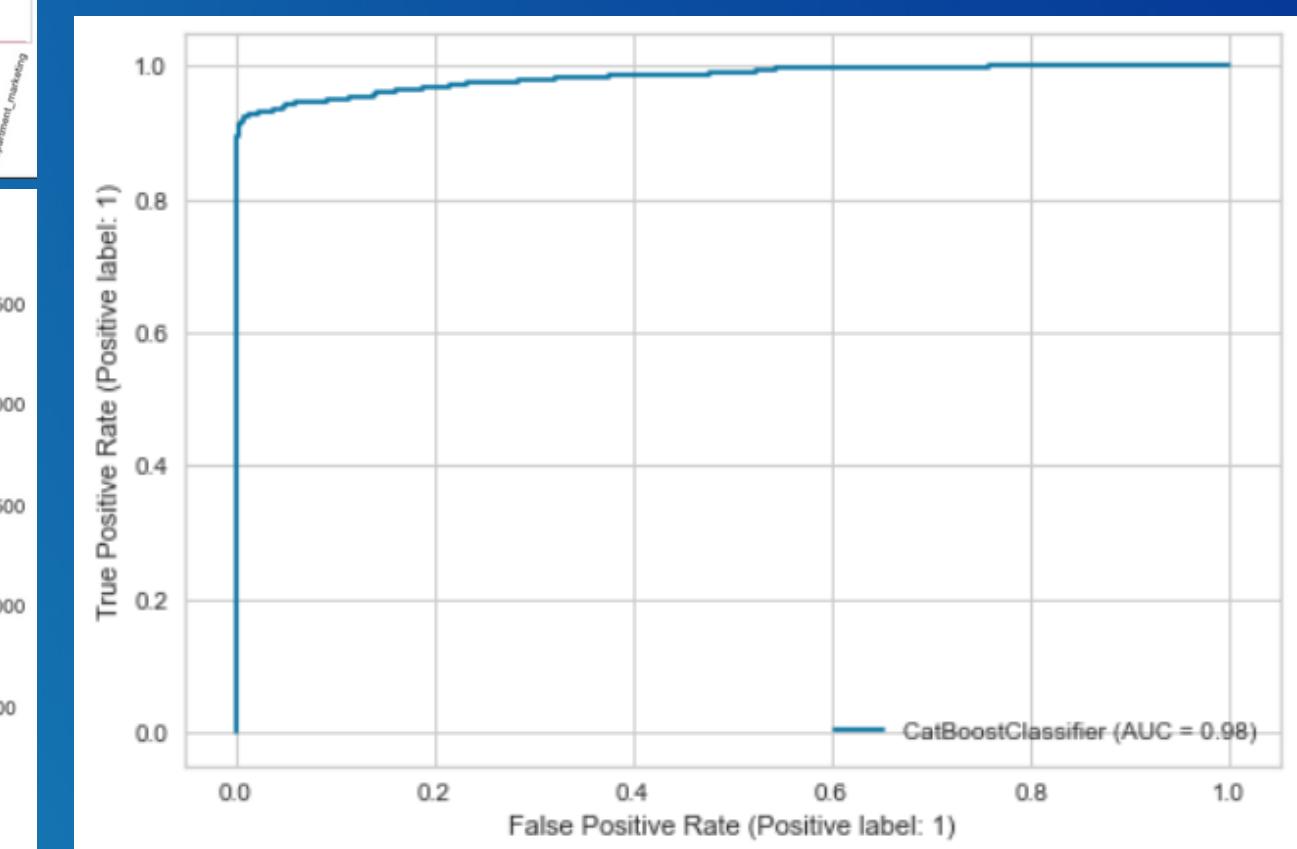
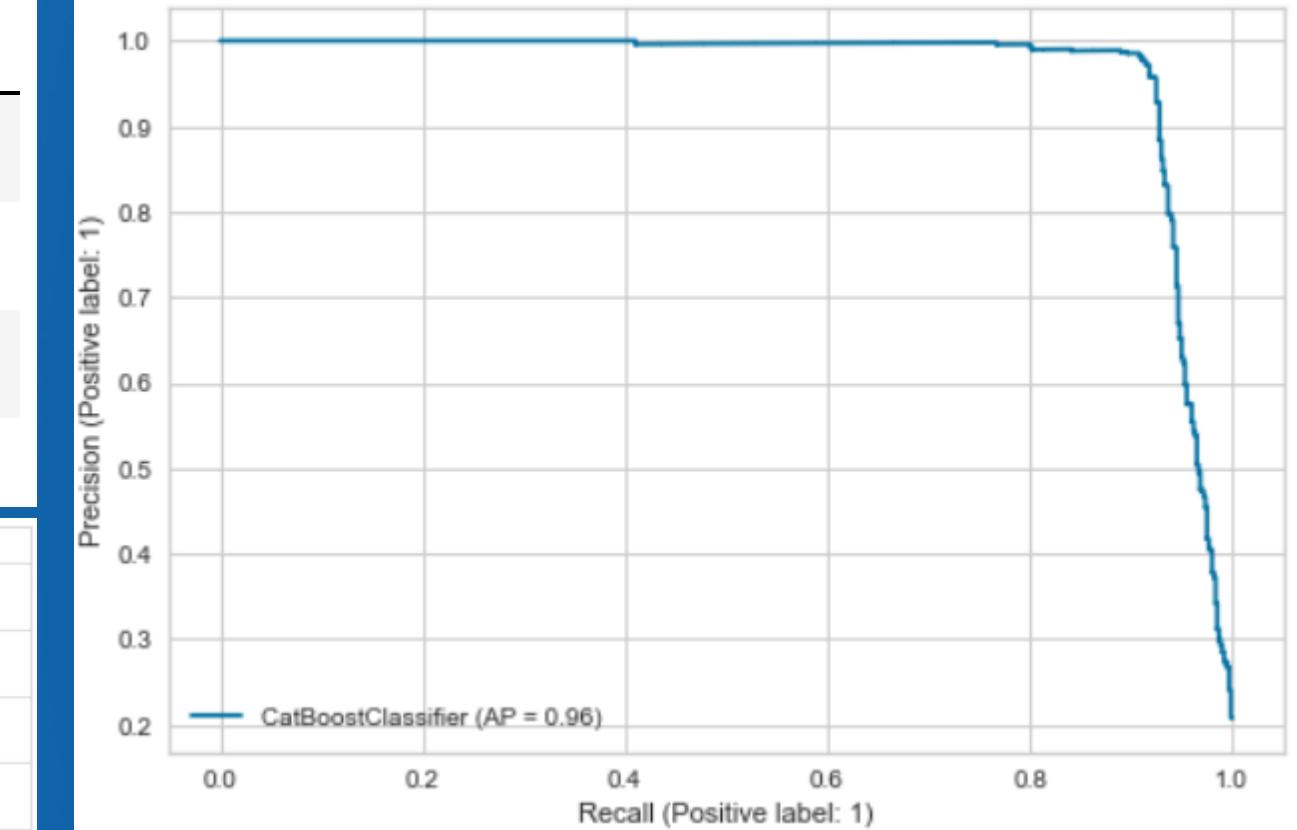
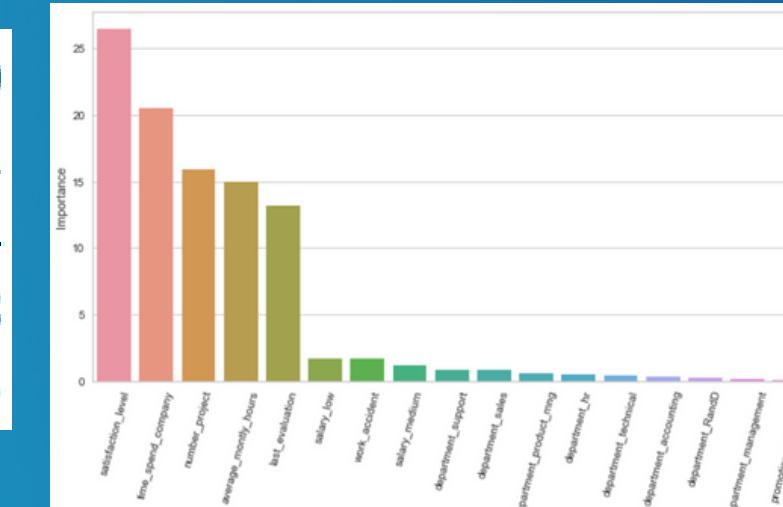
**test\_accuracy** 0.980  
**test\_precision** 0.952  
**test\_recall** 0.924  
**test\_f1** 0.938  
**test\_roc\_auc** 0.983

## CV

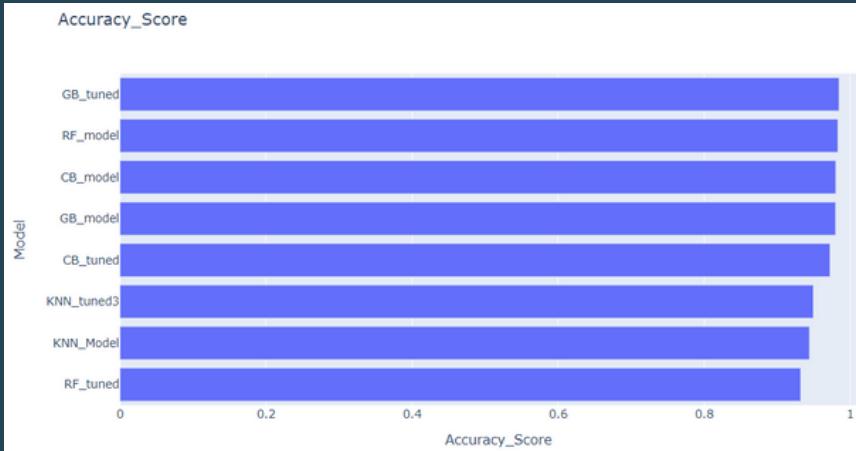


Best Parameters of GridSearchCV for Random Forest Model: {'depth': 4, 'l2\_leaf\_reg': 3, 'learning\_rate': 0.01}

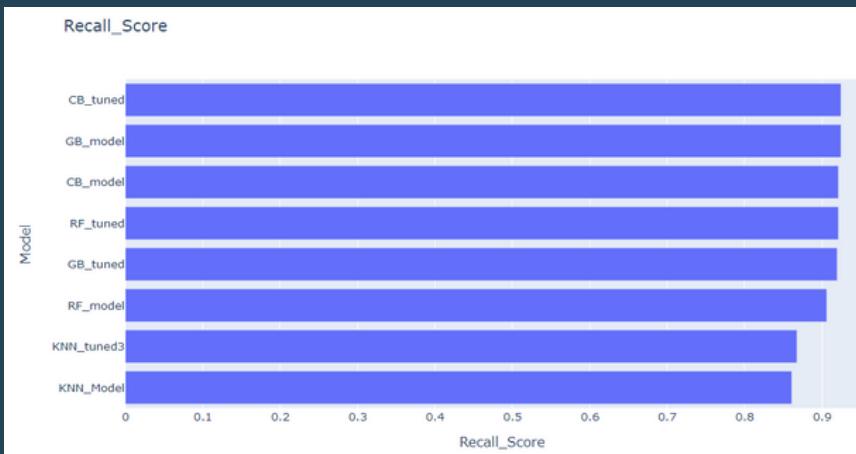
	train_set	test_set
<b>Accuracy</b>	0.977	0.972
<b>Precision</b>	0.923	0.909
<b>Recall</b>	0.940	0.925



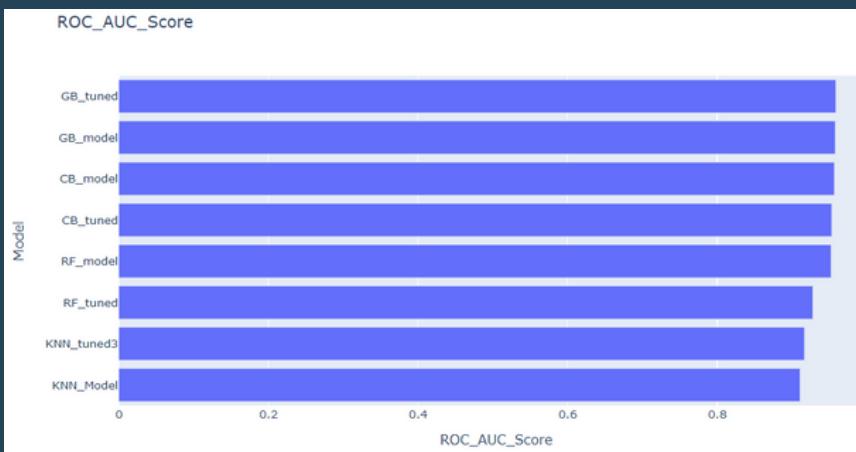
# MODEL EVALUATION



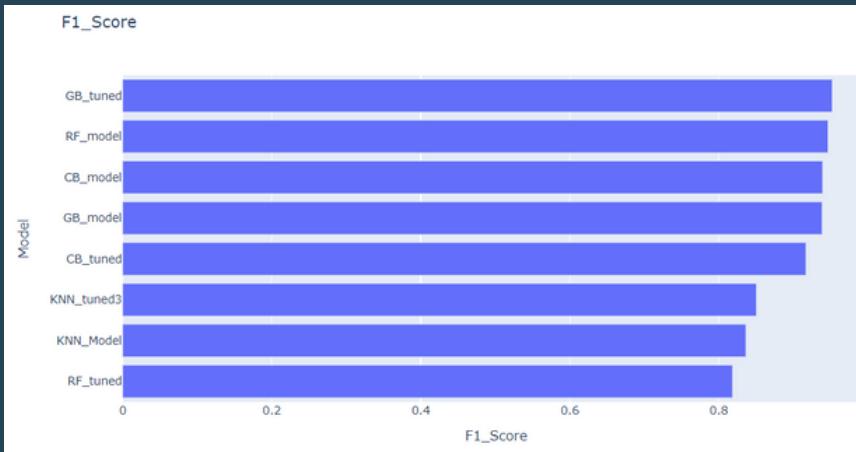
Gradient Boosting  
GridSearch CV



Catboost  
GridSearch CV



Gradient Boosting  
GridSearch CV



Gradient Boosting  
GridSearch CV

Actual	GB_Pred	KNN_Pred	RF_Pred	CB_Pred
8392	0	0	0	0
3803	0	0	0	0
8829	0	0	0	0
4734	0	0	0	0
3804	0	0	0	0
701	1	1	1	1
2475	0	0	0	1
1836	1	0	1	0
6925	0	0	0	1
10369	0	0	0	0

**True Positive**

Predicted:True  
Actual:True

**False Positive**

Predicted:True  
Actual:False

**True Negative**

Predicted:False  
Actual:False

**False Negative**

Predicted:False  
Actual:True

**Accuracy**

How often the model correct?

$$\frac{TP + TN}{Total}$$

**Recall**

When it actually is a positive case, how often is it correct?

$$\frac{TP}{Total\ Actual\ Pos.}$$

**Precision**

When prediction is positive, how often is it correct?

$$\frac{TP}{Total\ Predicted\ Pos.}$$

**F1-Score**

Harmonic mean of the precision and recall.

$$\frac{2 * Precision * Recall}{Precision + Recall}$$

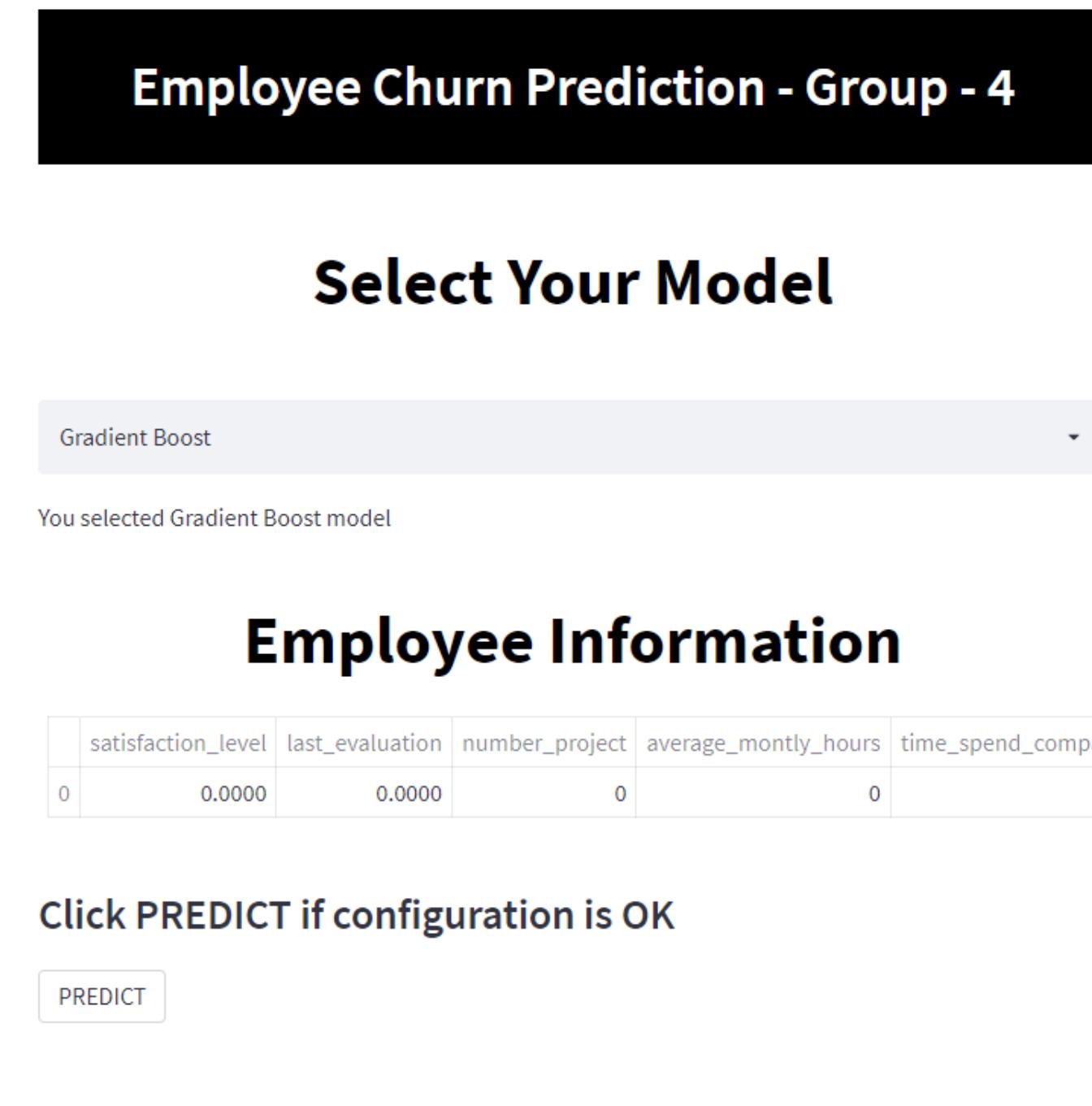
# MODEL DEPLOYMENT



EC2

<http://18.217.58.206:8501/>

Employee Information



The dashboard displays a title "Employee Churn Prediction - Group - 4" and a section "Select Your Model" with a dropdown menu set to "Gradient Boost". Below this, a message says "You selected Gradient Boost model". The main area is titled "Employee Information" and contains a table with employee data:

	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	promotion_last_5years	Departments	salary
0	0.0000	0.0000	0	0	0	1	1	1	RandD low

Click PREDICT if configuration is OK

PREDICT

Employee Information

Satisfaction Level

Last Evaluation

number\_project

average\_monthly\_hours

Time Spend in Company

# THANK YOU FOR YOUR PATIENCE...

molecules do not readily penetrate cell membrane. There are no specific transmembrane carriers for glucose because they are too polar to diffuse through membranes. The irreversible phosphorylation of glucose therefore, effectively traps the sugar as cytosolic glucose thus committing it to further metabolism in the several isozymes of the enzyme hexokinase that converts glucose to glucose 6-phosphate.

In most tissues, the phosphorylation of glucose is catalyzed by hexokinase, one of three regulatory enzymes of glycolysis (glucokinase and pyruvate kinase). Hexokinase has substrate specificity and is able to phosphorylate glucose in addition to glucose. Hexokinase is the rate-limiting enzyme in glycolysis. Its action product, glucose 6-phosphate, which is a key intermediate in further metabolism of this hexose phosphate, has a low  $K_m$  (and, therefore, a high affinity) for glucose. This permits the efficient phosphorylation of glucose even when tissue concentrations of glucose are low (Figure 8.13). Hexokinase, however, has a low  $V_{max}$  for glucose and, therefore, cannot phosphorylate more sugars than the cell can use.

In liver parenchymal cells and  $\beta$  cells of the pancreas (also called hexokinase II or hexokinase B), the

