I. Purpose

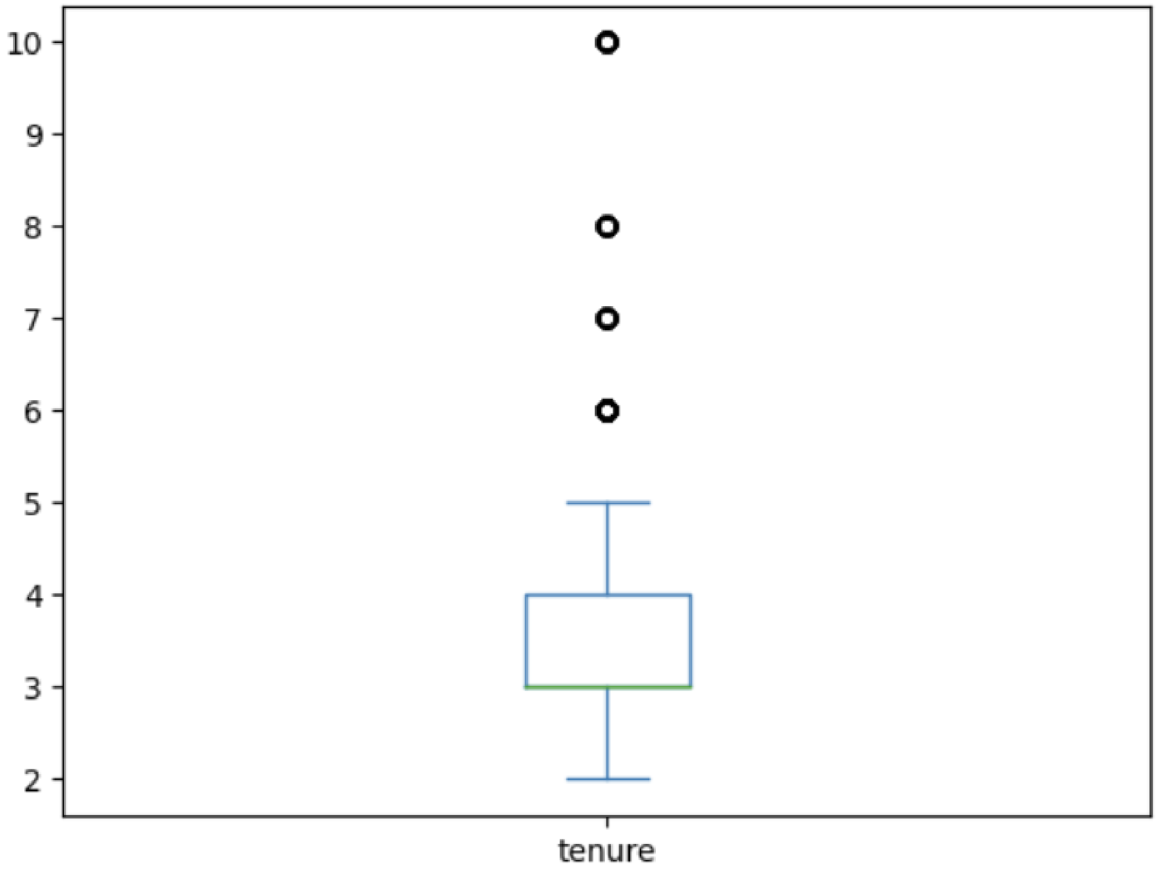
A good model can predict whether an employee will leave the company and find out why they leave, which will help the company better understand the problem and come up with solutions to increase retention and job satisfaction for current employees, while saving money and time training new employees.

II. Explore survey data

1. Size: 14999 samples, 9 features
2. Data type
   a. Continuous: satisfaction_level, last_performance_scr, num_projects, avg_monthly_hours, tenure
   b. Category: department, paid_rate
   c. Binary: had_work_accident, promoted_last_5_years, had_left_company

III. Cleaning survey data

1. Missing Value: None
2. Duplicate Value: 3008
3. Outlier: tenure

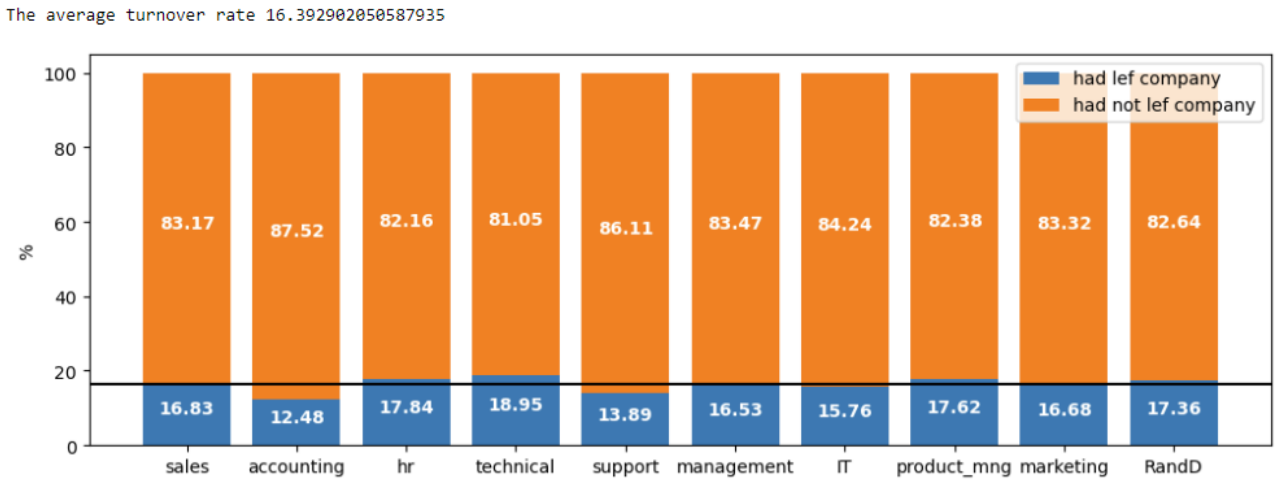


4. Sample size: 11167

a. Minority class (had_left_company = 1): 1882 (~ 83.15%)
b. Majority class (had_left_company = 0): 9285 (~ 16.85%) (~ 16.85%)

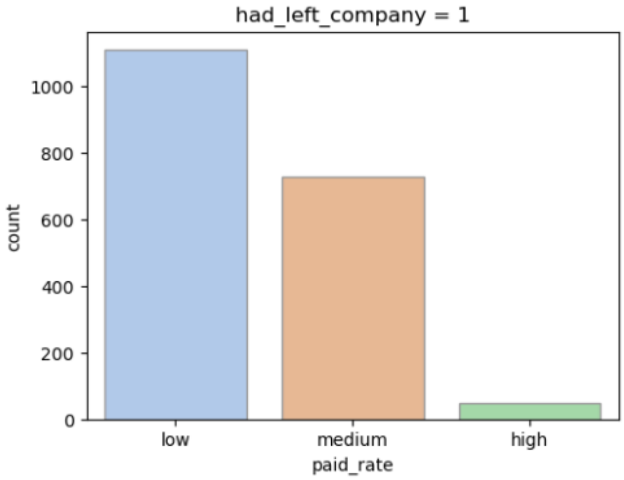
IV. Visual observation

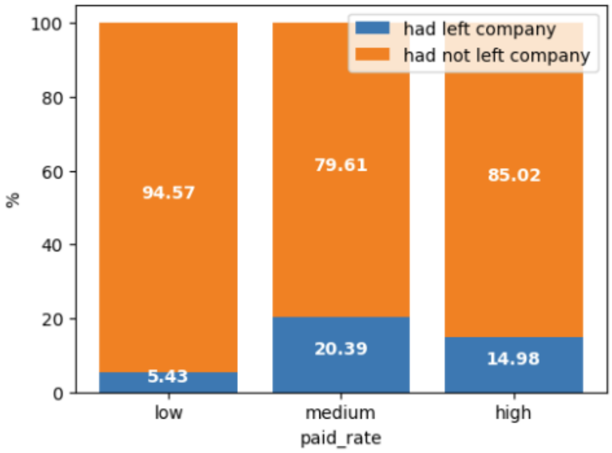
- The difference between the departments with the highest (engineering - 18.95%) and lowest (accounting - 12.48%) turnover rates is about 6.47%. With an average turnover rate of 16.4%, this is a large enough difference to suggest that departments may reflect differences in work environment, job stress, working conditions, or cultural factors between departments.



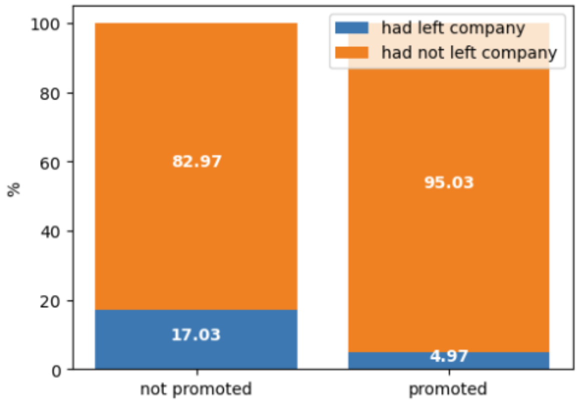
```
The average turnover rate 16.392902050587935
```

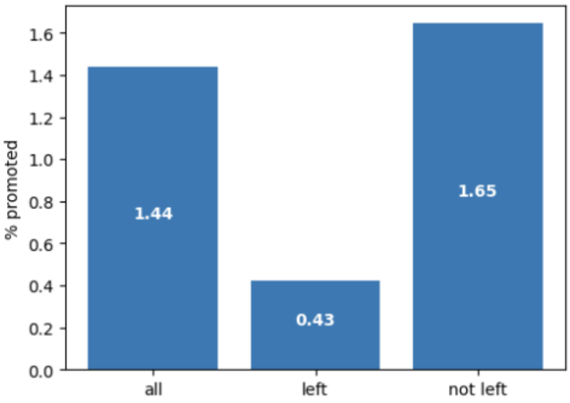
- Employees with higher salaries are less likely to leave a company due to satisfaction with compensation and benefits. At the same time, employees with lower salaries are more likely to stay because they have fewer other options or lower expectations. Average salaries are a more important factor in leaving a company, as this group has the highest turnover rate. However, low salaries also play a role, as the high number of low-paid employees leaving suggests that this is a significant problem.



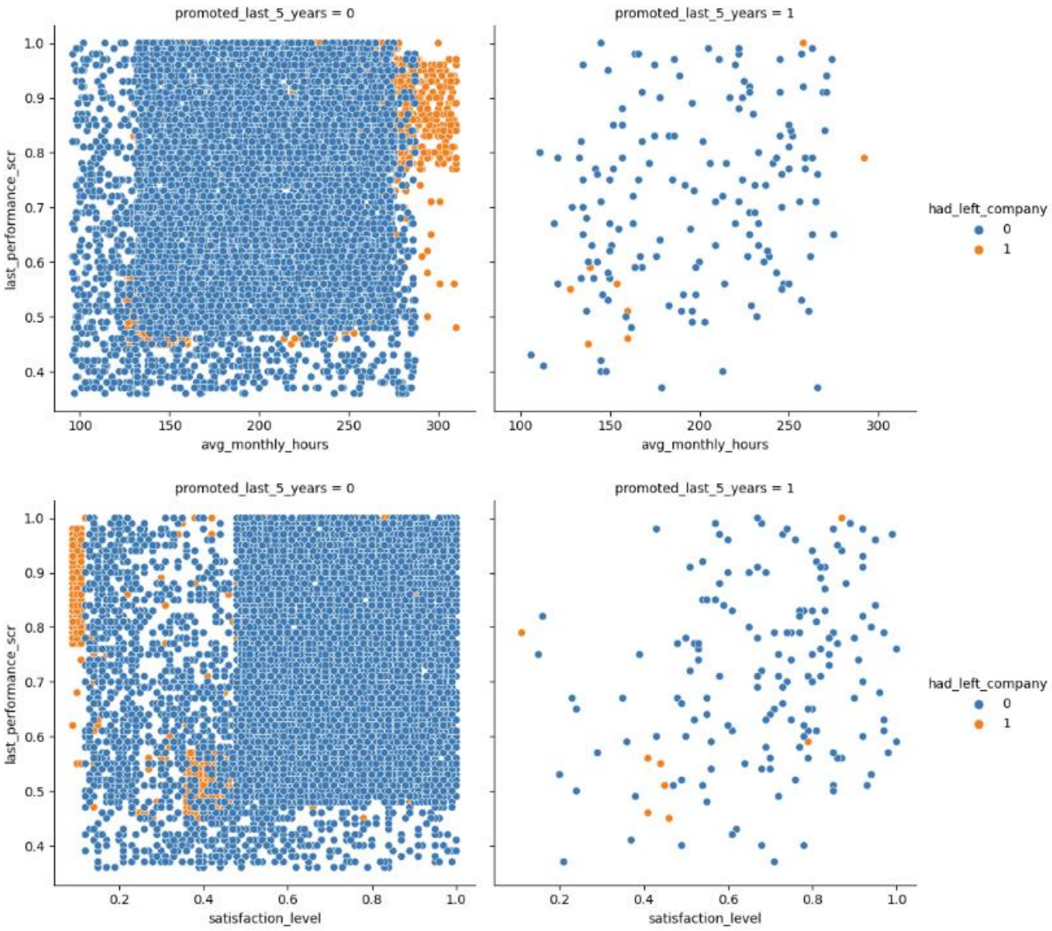


- The rate of leaving the company among the group of employees who were not promoted (17.03%) was much higher than the rate of leaving the company among the group who were promoted (4.97%). This suggests that not being promoted may be an important factor in making employees feel dissatisfied and decide to leave the company.

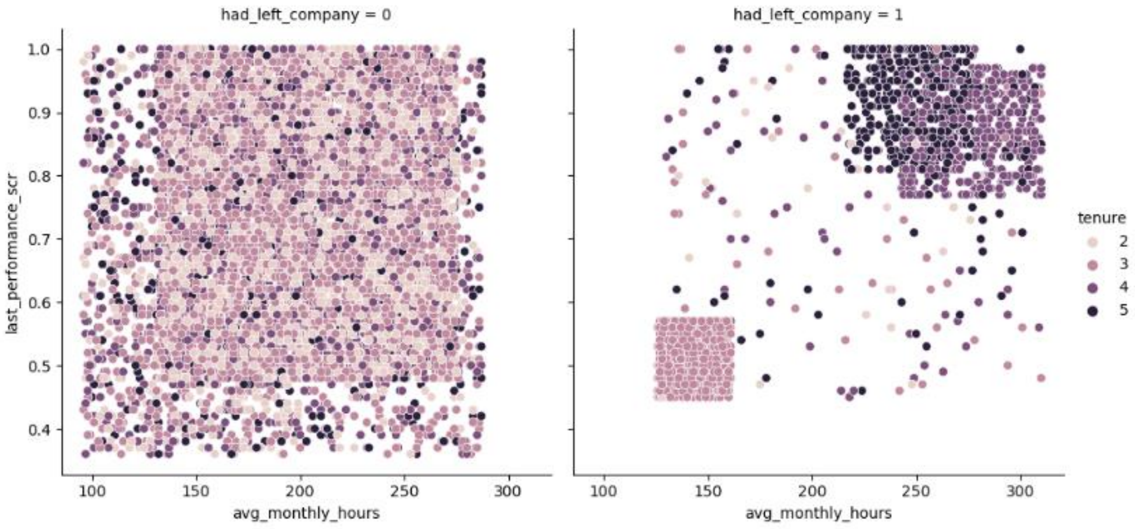


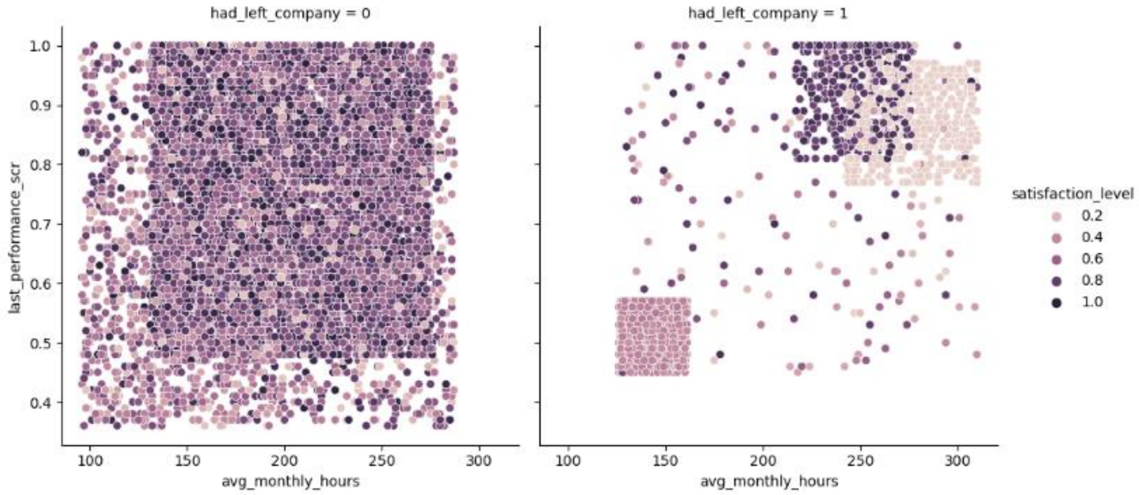


- High performers who are not promoted and feel dissatisfied tend to leave the company because they feel their efforts are not rewarded. Low-average performers also leave the company when they do not see opportunities for growth and satisfaction in their work



- High performance scores combined with very low satisfaction are major factors leading to turnover, despite relatively high productivity among employees working more than 250 hours per month. High performance but feeling stagnant or lacking growth opportunities may be a factor in those working more than 200 hours per month deciding to leave the company after five years.





V. Approach assessment

1. Problem

a. Multiple features can influence the decision to quit or leave the company.
b. The goal focuses on predicting that an employee will leave the company.
c. The feature importance in XGBoost only indicates the extent to which each feature contributes to the overall predictive performance of the model, not the change in the target variable when there is a unit change in the corresponding feature.
d. Multicollinearity and class imbalance affect the regression model.

2. Solution

a. Apply XGBoost to select the features that have the greatest impact on improving the model and prediction quality through the actual improvement in performance in the model's loss function when a feature is used to split the data.

b. Optimize GridSearchCV by Recall to minimize False Negative (when predicting that employees will stay with the company (had_left_company = 0), but in fact they had_left_company = 1).

c. The regression coefficient in the regression model measures the relationship between the feature variables and the target predictor variable, which can tell us the direction (positive or negative) and the degree of influence of each feature on the target variable.

d. Remove highly correlated features in the correlation matrix, and increase the number of samples of the minority class from 17% to 20% or 30% to improve the Recall and Precision of the minority class.

VI. Model building

1. Data split

- Training/Validation/Testing: 80/10/10
- One-hot encoding of category features

2. XGBoost

a. GridSearchCV: 10-fold cross-validation, refit = 'recall'

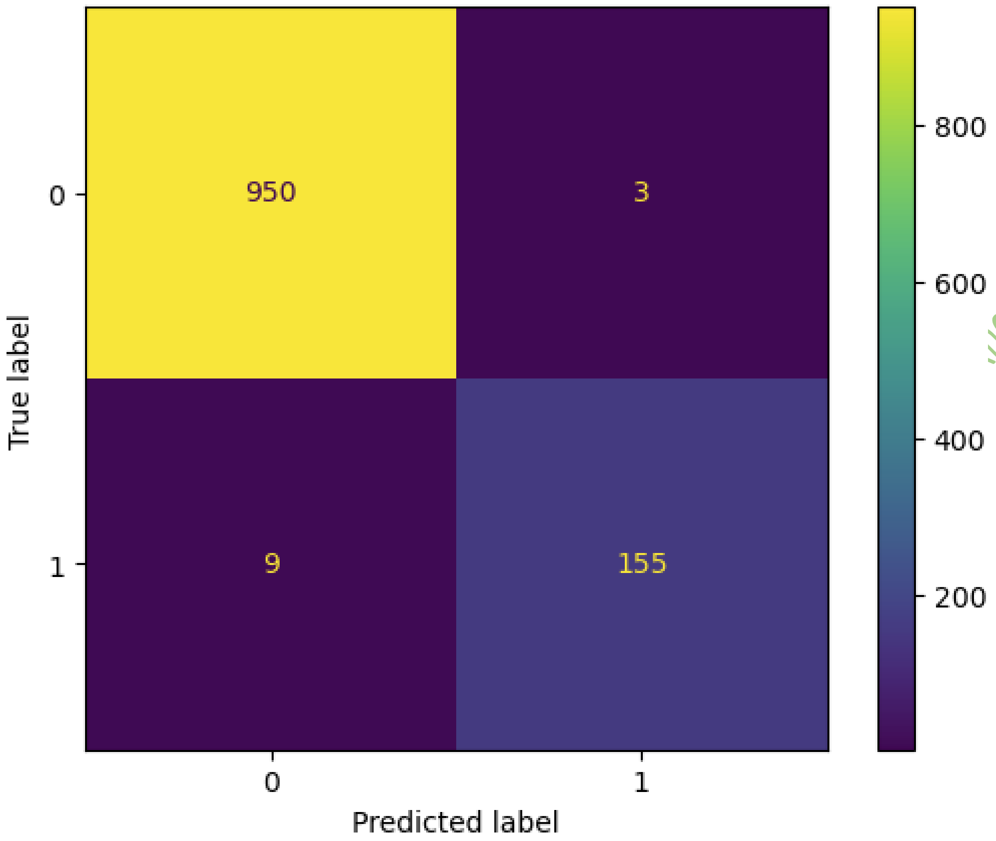
b. best_params_:

- {'learning_rate': 0.1,
- 'max_depth': 5,
- 'min_child_weight': 2,
- 'n_estimators': 500}

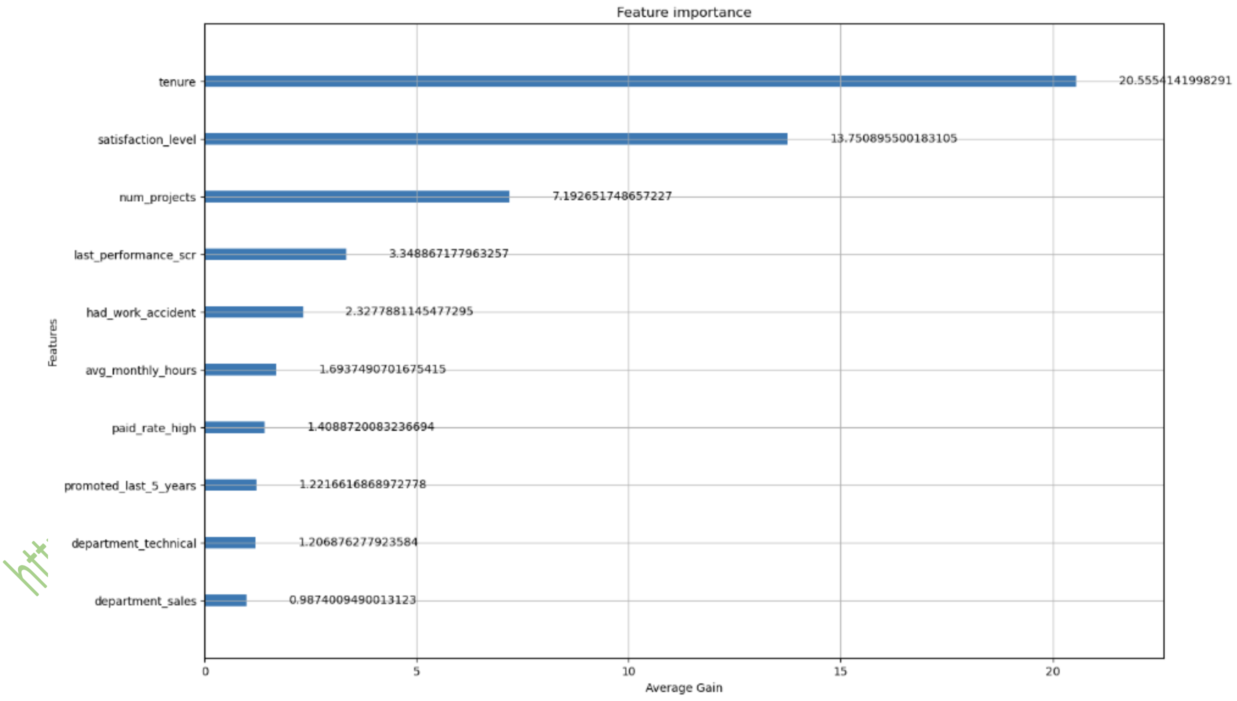
c. classification_report: Accuracy = 0.99

Class	Precision	Recall	F1-score
0	0.99	1.00	0.99
1	0.98	0.95	0.96

d. confusion_matrix:



e. plot_importance: importance_type='gain'



3. Logistic Regression

a. classification_report: Accuracy = 0.81

Class	Precision	Recall	F1-score
0	0.85	0.94	0.89
1	0.46	0.25	0.32

The basic Logistic Regression model has difficulty detecting cases belonging to the minority class (had_left_company = 1), as shown by the low Recall (0.25). This may result in missing many cases of employees leaving the company.

b. Chuẩn hóa dữ liệu: Accuracy = 0.81

Class	Precision	Recall	F1-score
0	0.85	0.94	0.89
1	0.46	0.23	0.31

Normalizing the data does not seem to bring much improvement to the model, with Recall even dropping slightly to 0.23. This may indicate that normalization is not very necessary in this problem.

c. Remove features with high linear relationship: Accuracy = 0.81

Class	Precision	Recall	F1-score
0	0.85	0.94	0.89
1	0.46	0.24	0.34

Removing highly linear features also does not make much difference, with the Recall for the minority class remaining low (0.24). This may indicate that the removed features do indeed contain important information for the model.

d. Select only important features: Accuracy = 0.81

Class	Precision	Recall	F1-score
0	0.85	0.94	0.89
1	0.46	0.26	0.34

Keeping only the important features gives a slight improvement in both Recall (0.26) of the minority class. This may be due to the model focusing on variables with higher statistical significance.

e. Increase the minority class ratio to 20%: Accuracy = 0.82

Class	Precision	Recall	F1-score
0	0.86	0.92	0.89
1	0.50	0.34	0.40

When increasing the minority class ratio to 20%, the Minority Class Recall improved significantly (0.34) compared to the baseline model, although the F1-Score was still not ideal.

f. Increase the minority class ratio to 25%: Accuracy = 0.84

Class	Precision	Recall	F1-score
0	0.90	0.91	0.90
1	0.57	0.54	0.56

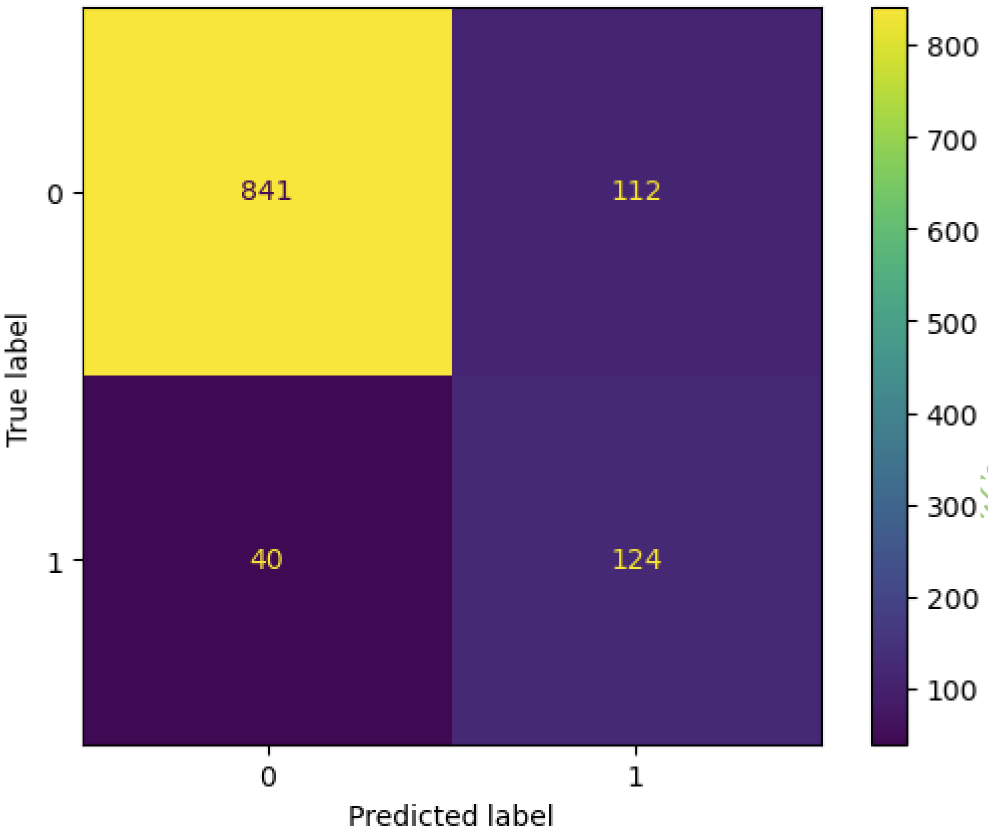
With the minority class ratio of 25%, both the Precision and Recall of the minority class increased sharply, reaching almost balanced levels (0.57 and 0.54), indicating that the model detected the minority class instances better.

g. Increase the minority class ratio to 30%: Accuracy = 0.84

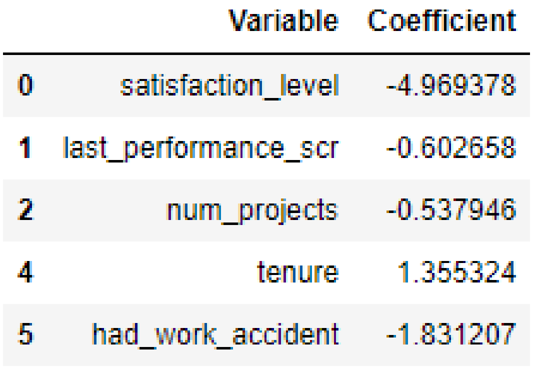
Class	Precision	Recall	F1-score
0	0.93	0.89	0.91
1	0.58	0.69	0.63

With an upsampling ratio of 30%, the model shows significant improvement with Recall reaching 0.69 and F1-Score reaching 0.63, which are the best results in the experiments, indicating that this upsampling ratio is optimal in improving model performance on the minority class

h. confusion_matrix



i. Regression coefficient



	Variable	Coefficient
0	satisfaction_level	-4.969378
1	last_performance_scr	-0.602658
2	num_projects	-0.537946
4	tenure	1.355324
5	had_work_accident	-1.831207

VII. Evaluation of results

1. Tenure has a positive regression coefficient (+1.355): This shows that as an employee's years of service at a company increase, their likelihood of leaving also increases. Employees who have been working for a long time may feel like they want a change, looking for new opportunities, or feeling tired of their current job. They may reach a stage in their career where they want a new challenge or are looking for a better work-life balance.

2. Satisfaction Level has a negative regression coefficient (-4.969): This is a very large negative coefficient, indicating that as employee satisfaction increases, their

likelihood of leaving decreases sharply. This makes sense, as employees who are satisfied with their jobs tend to stay with the company longer. They feel respected, and motivated and find their work meaningful.

3. The number of Projects has a negative regression coefficient (-0.537): The negative coefficient indicates that as the number of projects an employee participates in increases, the likelihood of them leaving the company decreases. Employees who participate in many projects may feel that they have an important role in the company, are challenged, and have the opportunity to develop their skills. This may increase their commitment to the company, thereby reducing the likelihood of leaving the company.

4. Last Performance Score has a negative regression coefficient (-0.603): This indicates that as an employee's recent performance score increases, the likelihood of them leaving the company decreases. Employees with high performance may feel recognized and appreciated by the company. Being appreciated can bring about job satisfaction and a sense of security about their job, thereby reducing the intention to leave the company.

5. Had Work Accident has a negative regression coefficient (-1.831): This actually means that employees who have had work accidents are less likely to leave their jobs than employees who have not had work accidents. It is possible that the company provided good support and care to employees after they had a work accident, leading to them feeling cared for and more committed to the company.

VIII. Suggested improvements

❖ Long-term employees may feel the urge to change jobs or look for new opportunities when they do not see much challenge or progression in their careers. By providing career development opportunities, such as promotions, training in new skills, or internal job rotations, you can help reduce boredom and burnout, while also providing employees with valuable experience and knowledge to stay and grow with the company.

❖ Provide opportunities for employees to participate in a variety of projects and ensure that they are given challenging and meaningful tasks. This can be coupled with recognition and achievement programs to increase employee satisfaction. Employee satisfaction is closely linked to whether or not they stay with a company. When employees are involved in a variety of projects, they feel important and have the opportunity to develop their skills, which reduces the likelihood of them leaving the company.

- ❖ Establish a fair and transparent performance appraisal system and use the results to make decisions about promotions, raises, or rewards. Conduct regular surveys to measure employee satisfaction and identify sources of dissatisfaction. This will create recognition and motivation for high-performing employees.

- ❖ Companies should provide health support programs, insurance, and reasonable rest periods for employees who have work-related injuries. At the same time, create policies that encourage return to work after an accident, helping them feel cared for and protected. Employees who have work-related injuries are less likely to leave when they receive good support from the company.