I. Purpose

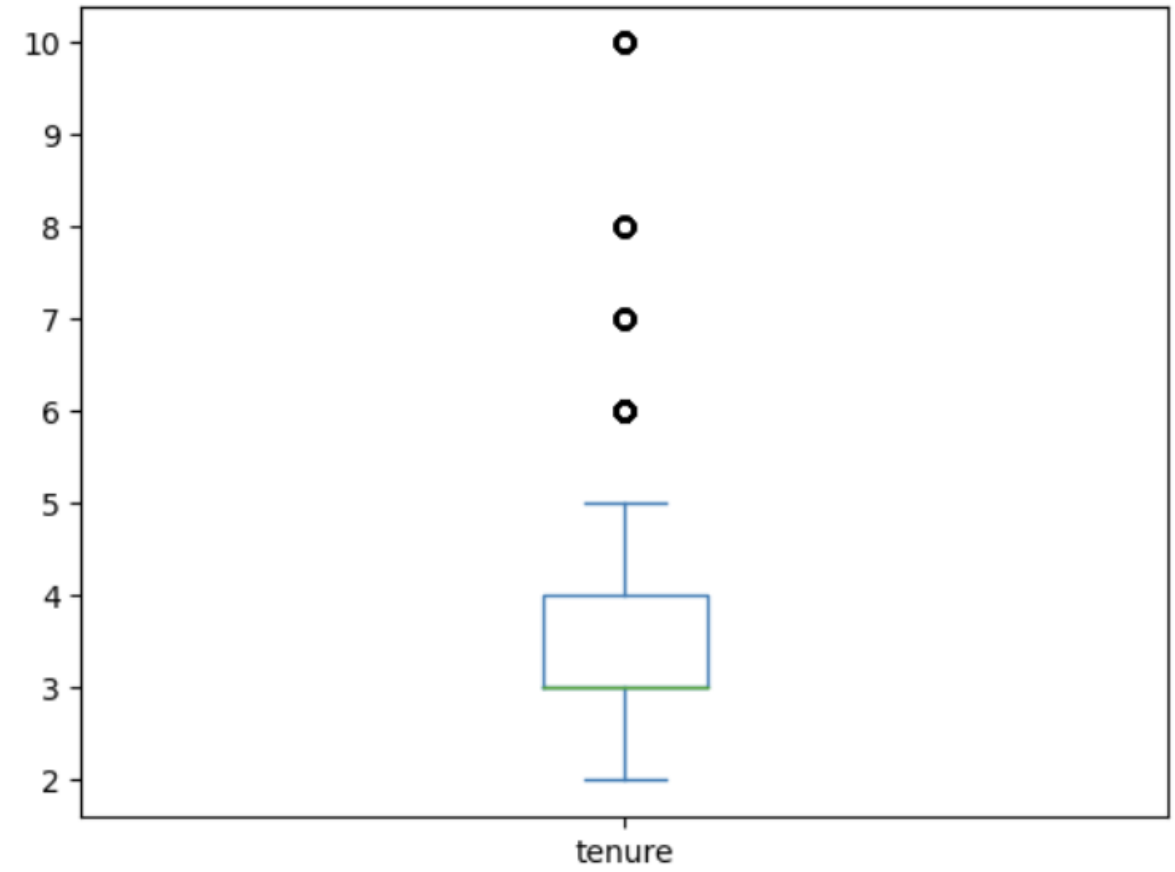
A good model can predict whether an employee will leave the company and find out why they leave, which will help the company better understand the problem and come up with solutions to increase retention and job satisfaction for current employees, while saving money and time training new employees.

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

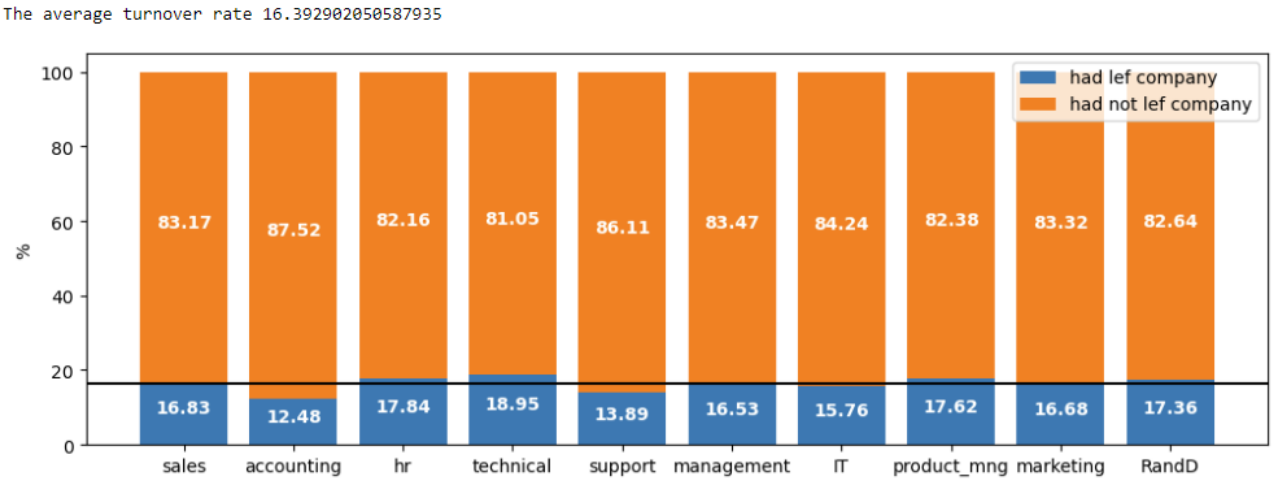


4. Sample size: 11167

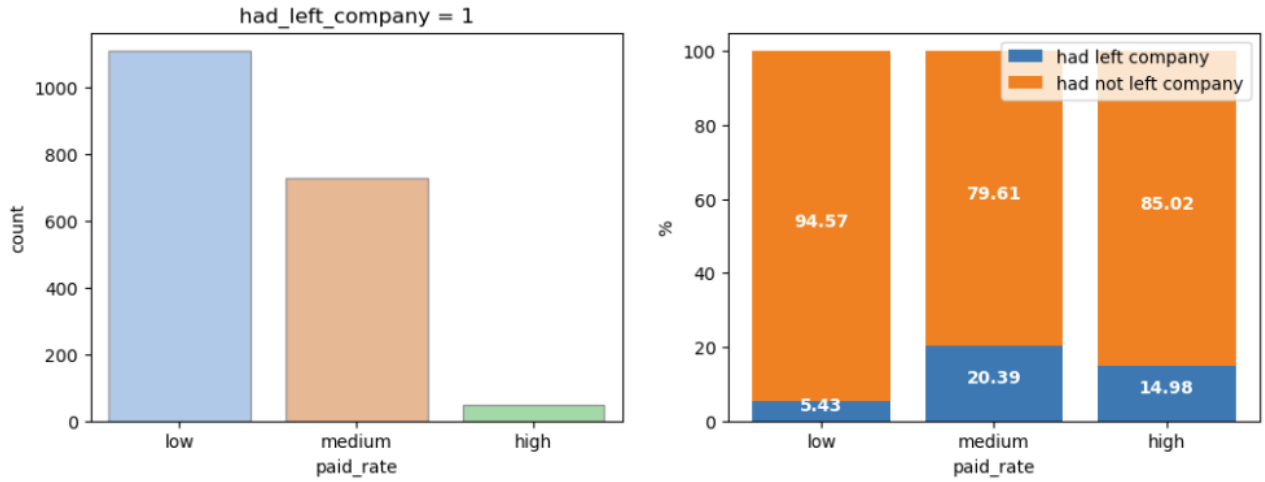
a. Minority class (had_left_company = 1): 1882 (~ 83.15%)
b. Majority class (had_left_company = 0): 9285 (~ 16.85%) (~ 16.85%)

IV. Visual observation

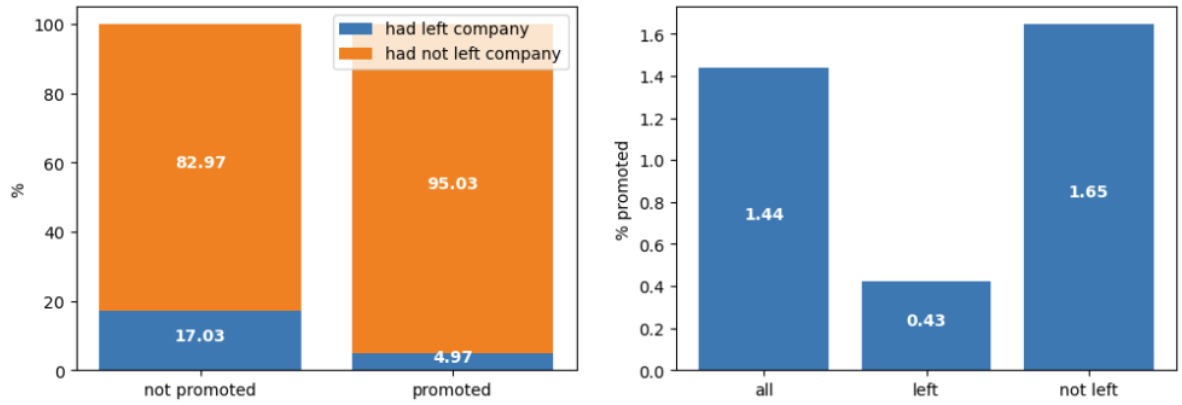
- The difference between the departments with the highest (engineering - 18.95%) and lowest (accounting - 12.48%) turnover rates is about 6.47%. With an average turnover rate of 16.4%, this is a large enough difference to suggest that departments may reflect differences in work environment, job stress,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



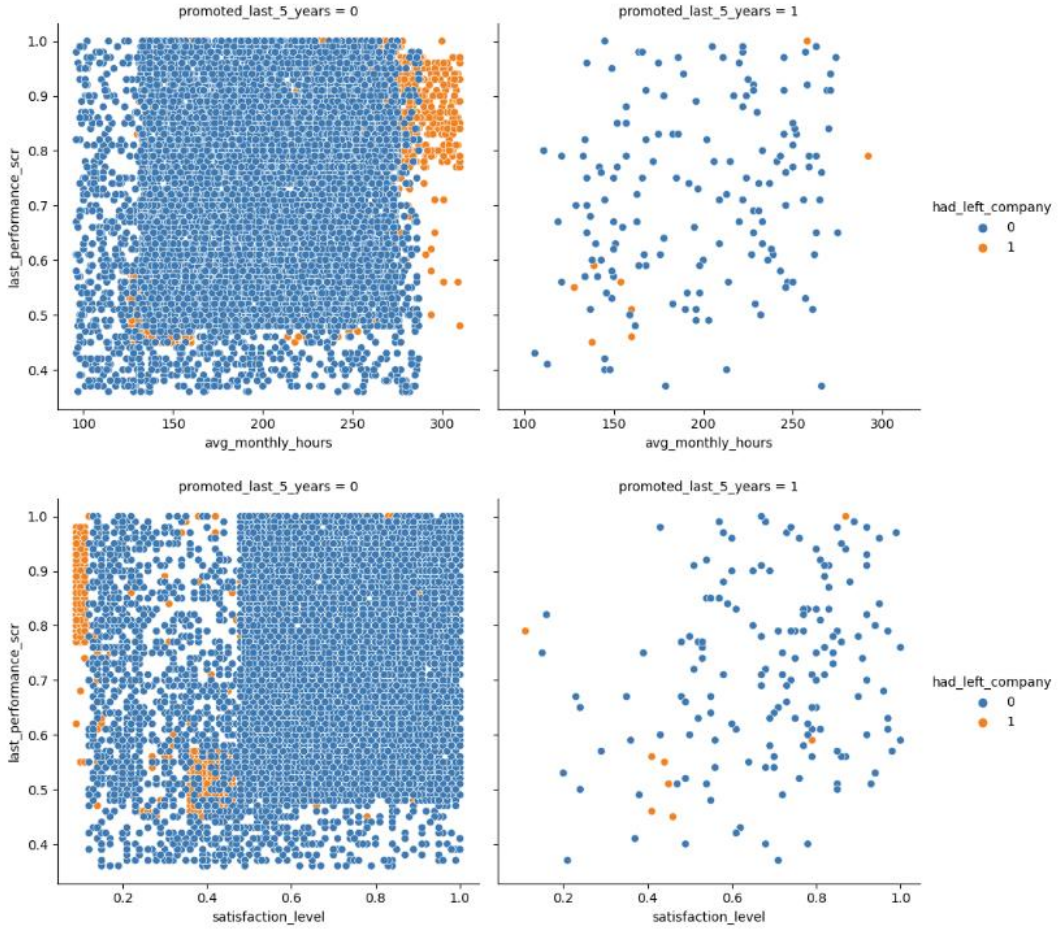
- Employees with higher salaries are less likely to leave a company due to satisfaction with compensation and benefits. At the same time, employees with lower salaries are more likely to stay because they have fewer other options or lower expectations. Average salaries are a more important factor in 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.



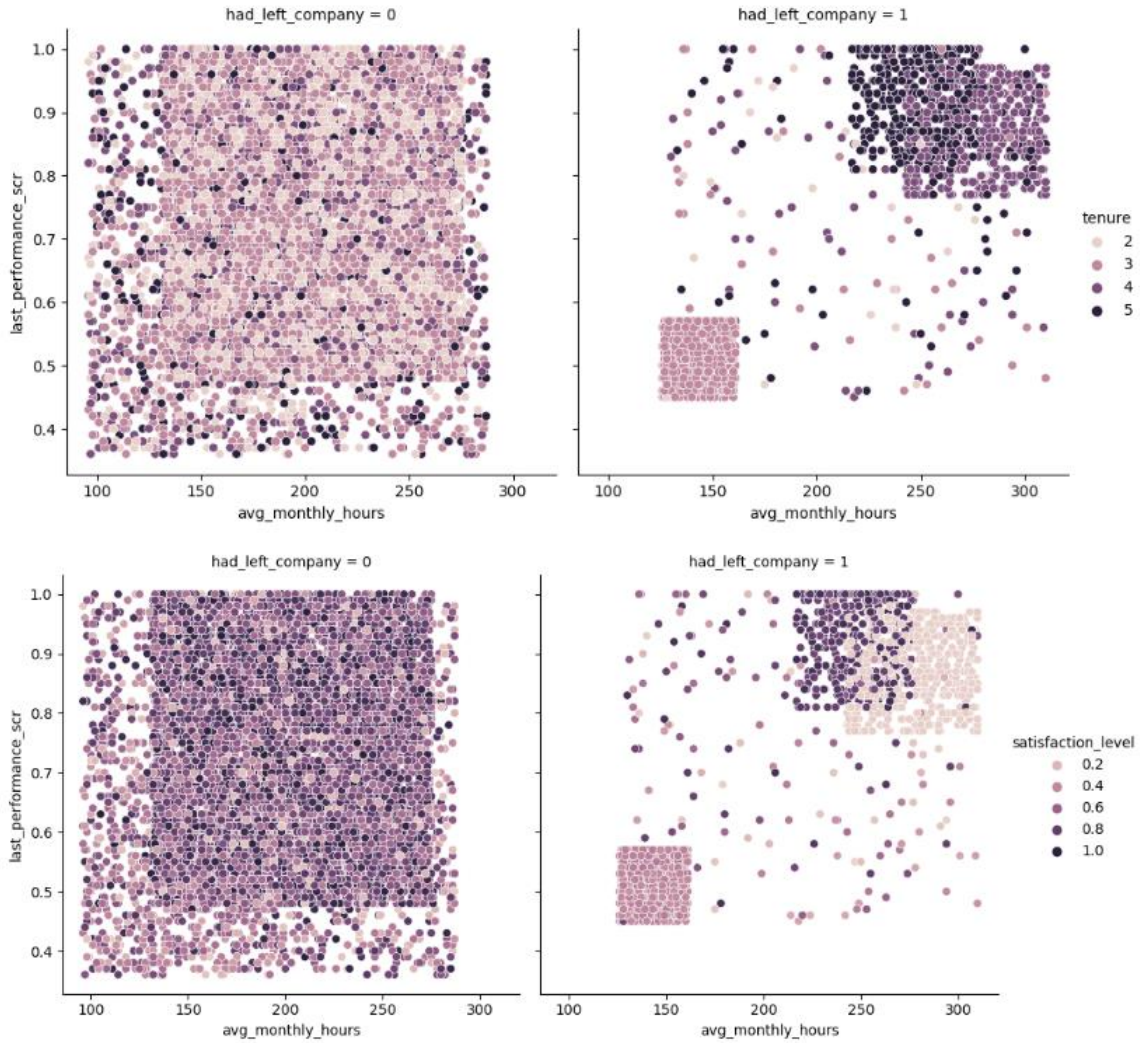
- The rate of leaving the company among the group of employees who were not promoted (17.03%) was much higher than the rate of leaving the company among the group who were promoted (4.97%). This suggests that not being promoted may be an important factor in making employees feel dissatisfied and decide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

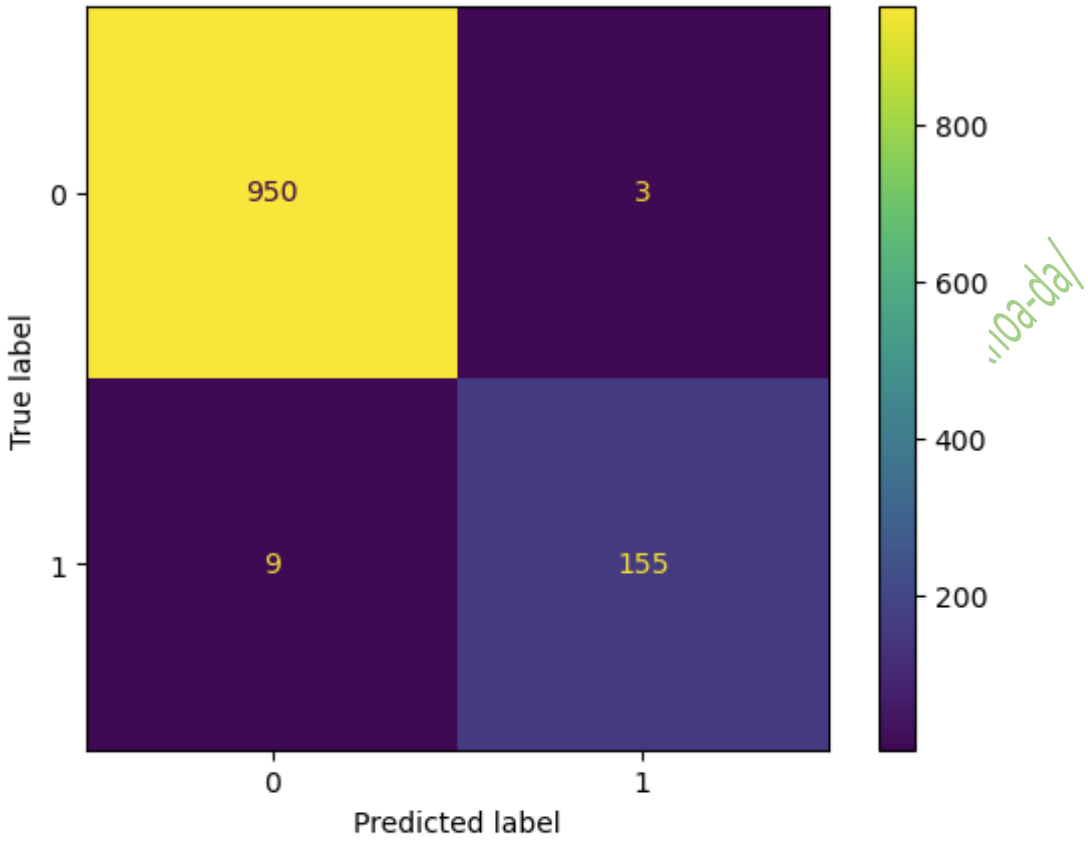
VI. Model building

1. Data split
    - Training/Validation/Testing: 80/10/10
    - One-hot encoding of category features

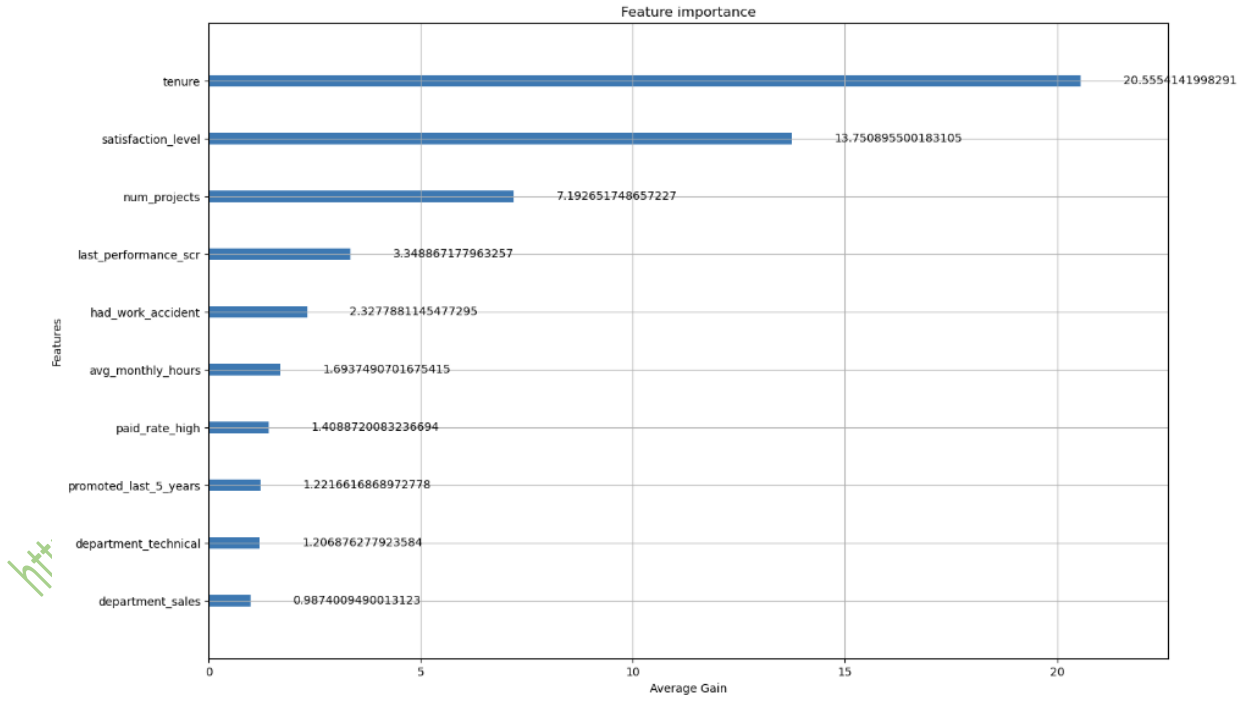
2. XGBoost
    a. GridSearchCV: 10-fold cross-validation, refit = 'recall'
    b. best_params_:
        - {'learning_rate': 0.1,
        - 'max_depth': 5,
        - 'min_child_weight': 2,
        - 'n_estimators': 500}
    c. classification_report: Accuracy = 0.99

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

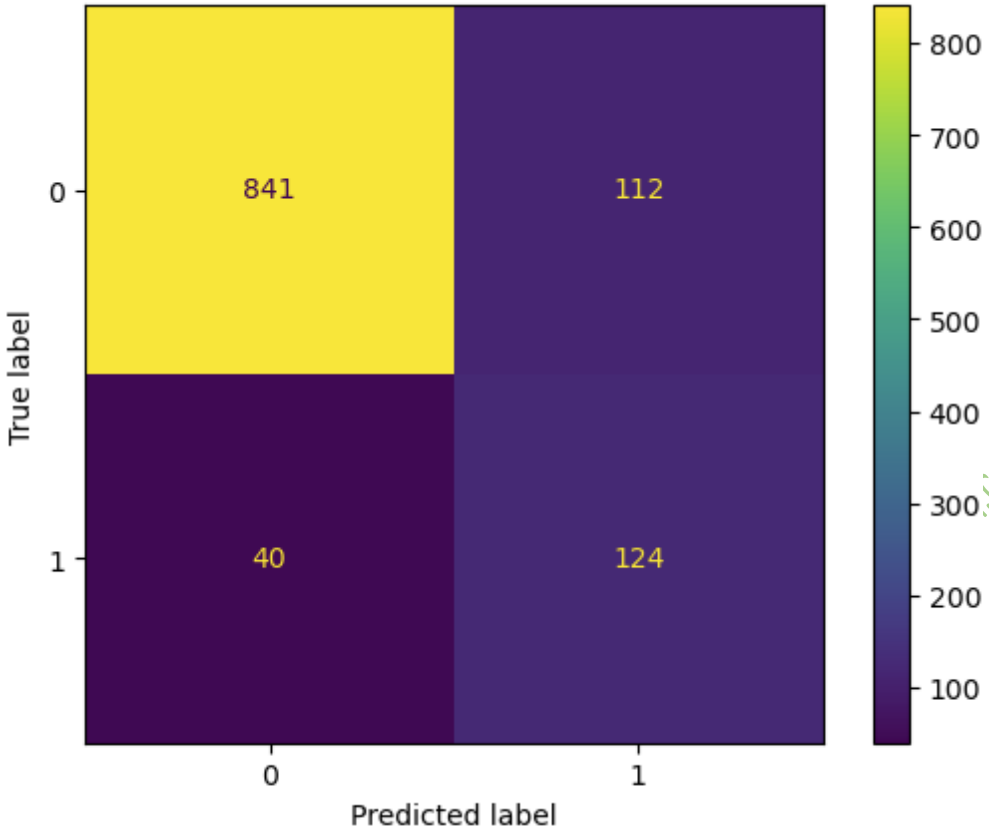
With the minority class ratio of 25%, both the Precision and Recall of the minority class increased sharply, reaching almost balanced levels (0.57 and 0.54), indicating that the model detected the minority class instances better.

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



4. Hypothesis testing

a. Null Hypothesis

(H0) = There is NO difference in the likelihood of leaving the job between employee groups (when controlling for tenure, satisfaction, and the number of projects). Essentially, the odds ratio for the employee group variable equals 1, meaning no effect.

b. Alternative Hypothesis

(H1)=There IS a difference in the likelihood of leaving the job between employee groups(after accounting for the other variables). The odds ratio for the employee group variable is significantly different from 1, indicating either a higher or lower likelihood of turnover.

c. Statistical Technique: logistic regression

d. Dependent variable: *'had_left_company'*

e. Independent variables

- *'had_work_accident'*
- *'paid_rate_low'*

f. Independent variables

- *'tenure'*

- *'satisfaction_level'*
- *'num_projects'*
- *'last_performance_scr'*
- *'avg_monthly_hours'*

g. P-value

```
                           Logit Regression Results
==============================================================================
Dep. Variable:       had_left_company   No. Observations:                11167
Model:                         Logit   Df Residuals:                    11159
Method:                          MLE   Df Model:                            7
Date:                Sat, 24 Aug 2024   Pseudo R-squ.:                 0.2924
Time:                        14:06:55   Log-Likelihood:                -3583.9
converged:                      True   LL-Null:                       -5064.8
Covariance Type:            nonrobust   LLR p-value:                     0.000
==========================================================================================
                             coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------
Intercept                 -1.7011      0.177     -9.631      0.000      -2.047      -1.355
had_work_accident         -1.4953      0.120    -12.460      0.000      -1.731      -1.260
paid_rate_low              0.5247      0.061      8.647      0.000       0.406       0.644
tenure                     1.0607      0.039     27.041      0.000       0.984       1.138
satisfaction_level        -4.5562      0.136    -33.445      0.000      -4.823      -4.289
num_projects              -0.4761      0.030    -15.935      0.000      -0.535      -0.418
last_performance_scr      -0.0167      0.203     -0.082      0.935      -0.414       0.381
avg_monthly_hours          0.0035      0.001      5.028      0.000       0.002       0.005
==========================================================================================
Optimization terminated successfully.
         Current function value: 0.320936
         Iterations 7
```

- The p-value for *had_work_accident* is 0.000: Very small, lower than the 0.05 significance level, so we can reject H0. This shows that there is a statistically significant relationship between having a work accident and the likelihood of leaving the job.
- The p-value for *paid_rate_low* is 0.000: Very small, lower than the 0.05 significance level, so we can reject H0. This indicates that low salary has a significant effect on the likelihood of leaving the job.
- P-value for *tenure* is 0.000: Very small, lower than the significance level of 0.05, so we can reject H0. Number of years of service has a significant effect on the likelihood of leaving the job.
- P-value for *satisfaction_level* is 0.000: Very small, lower than the significance level of 0.05, so we can reject H0. Satisfaction level has a significant effect on the likelihood of leaving the job.
- P-value for *num_projects* is 0.000: Very small, lower than the significance level of 0.05, so we can reject H0. Number of projects has a significant effect on the likelihood of leaving the job.

- P-value for *last_performance_scr* is 0.935: Greater than 0.05, so we cannot reject H0. This shows that the recent performance score does not have a significant effect on the likelihood of leaving.
- The P-value for *avg_monthly_hours* is 0.000: Very small, lower than the 0.05 significance level, so we can reject H0. The average number of hours worked per month has a significant effect on the likelihood of leaving.

In summary, with these results, we can reject the Null Hypothesis (H0) for all variables except "last_performance_scr". This shows that there is a statistically significant difference in the likelihood of leaving between groups of employees for most of the controlled variables.

h. Interpret the coefficients as odds ratios

	Feature	Coefficient	Odds Ratio
1	had_work_accident	-1.504393	0.222152
2	paid_rate_low	0.527413	1.694543
3	tenure	1.072799	2.923551
4	satisfaction_level	-4.494139	0.011174
5	num_projects	-0.404405	0.667374

VII. Evaluation of results

1. *'tenure'* has a regression coefficient of 1.072799, resulting in an odds ratio of 2.923551. This means that for each additional year of employment, an employee is 2.92 times more likely to leave the company.

2. *'satisfaction_level'* has a regression coefficient of -4.494139, resulting in an odds ratio of 0.011174. This means that as employee satisfaction increases, the likelihood of leaving the company decreases significantly, by about 98.9%.

3. '*num_projects'* has a regression coefficient of -0.404405, resulting in an odds ratio of 0.667374. This suggests that a higher number of projects can reduce the likelihood of leaving the company by about 33.3%.

4. '*had_work_accident'* has a regression coefficient of -1.504393, resulting in an odds ratio of 0.222152. This means that employees who have had a work accident are about 77.8% less likely to leave the company than those who have not had a work accident.

5. *'paid_rate_low'* has a regression coefficient of 0.527413, resulting in an odds ratio of 1.694543. This means that employees with low salaries are about 69% more likely to leave the company than employees with higher salaries.

VIII. Suggested improvements

1. Ensure competitive salaries: Compare the company's current salary with the market and adjust to ensure competitiveness. This is especially important for low-wage employees, as the analysis has shown that they are more likely to leave the company.

2. Provide opportunities for employees to participate in many projects: Encourage employees to participate in many projects to help them feel challenged and have the opportunity to develop their skills. As the analysis has shown, the number of projects an employee participates in is associated with a lower rate of leaving the company.

4. Support employees after a work accident: Provide recovery support programs for employees who have had a work accident, including psychological counseling, medical support, and retraining programs if necessary. The analysis shows that having a work accident can be related to good support from the company. Maintaining or even improving post-work accident support policies can help retain employees better.

5. Develop a long-term commitment program: For employees with high seniority, the company can develop special programs such as additional vacation, seniority bonuses, and promotion opportunities to retain them. The analysis results show that the longer employees work at the company, the more likely they are to want to leave, possibly because they feel like they want a change, are looking for new opportunities, or are tired of their current job.