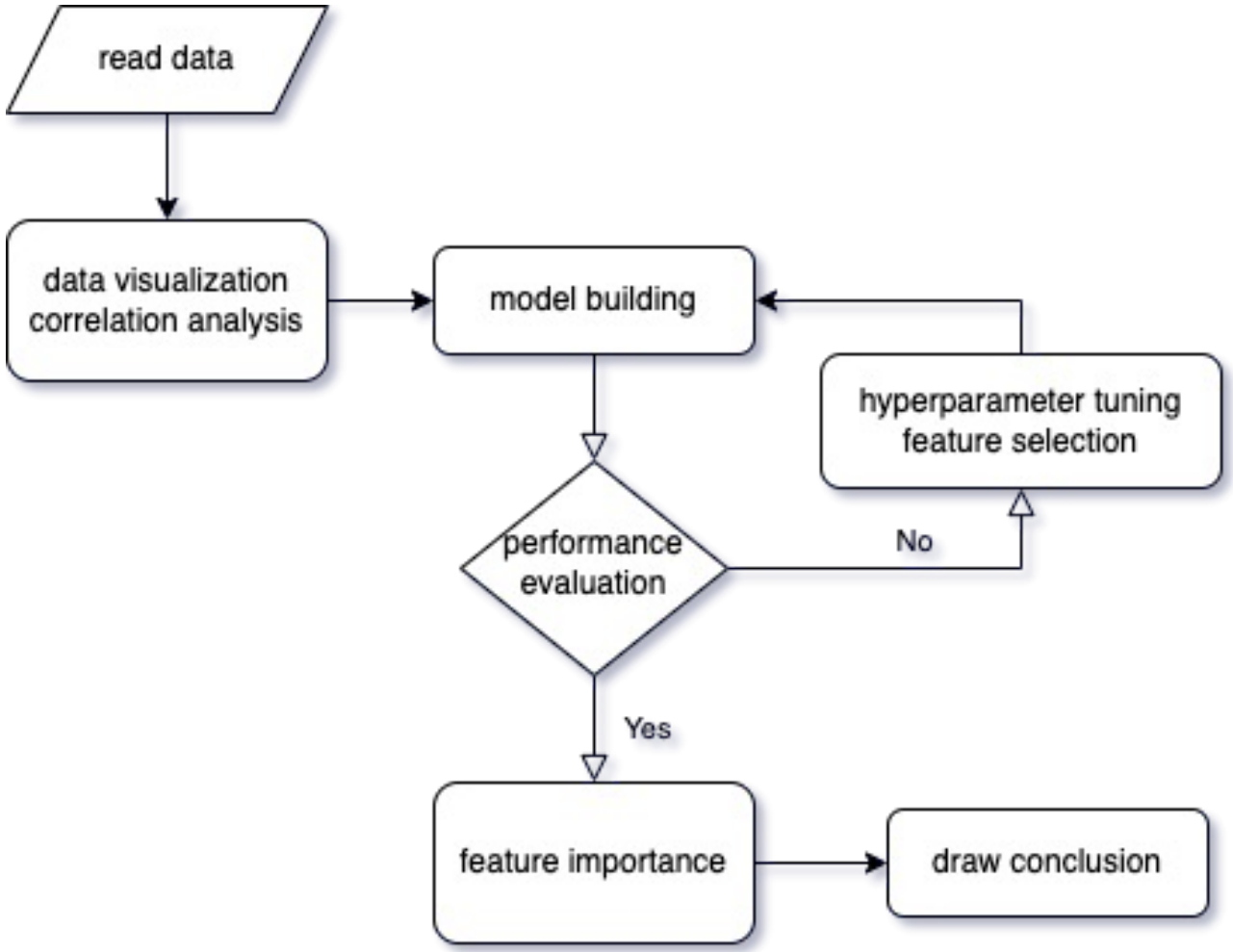
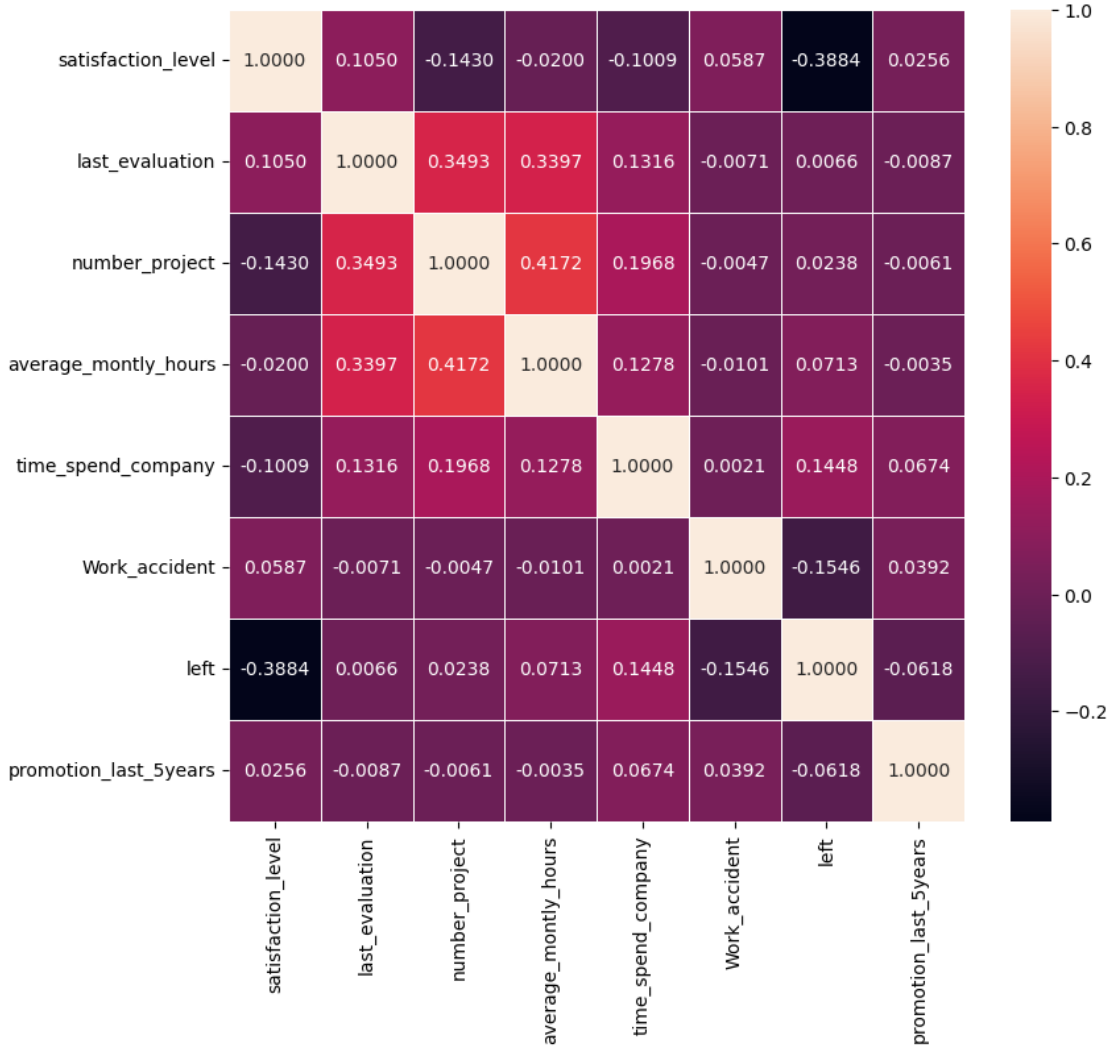
Data Science Project

- Understand why employees are leaving the company
- Who will be the next ones to leave
- Find an action plan to tackle this problem

Pipeline

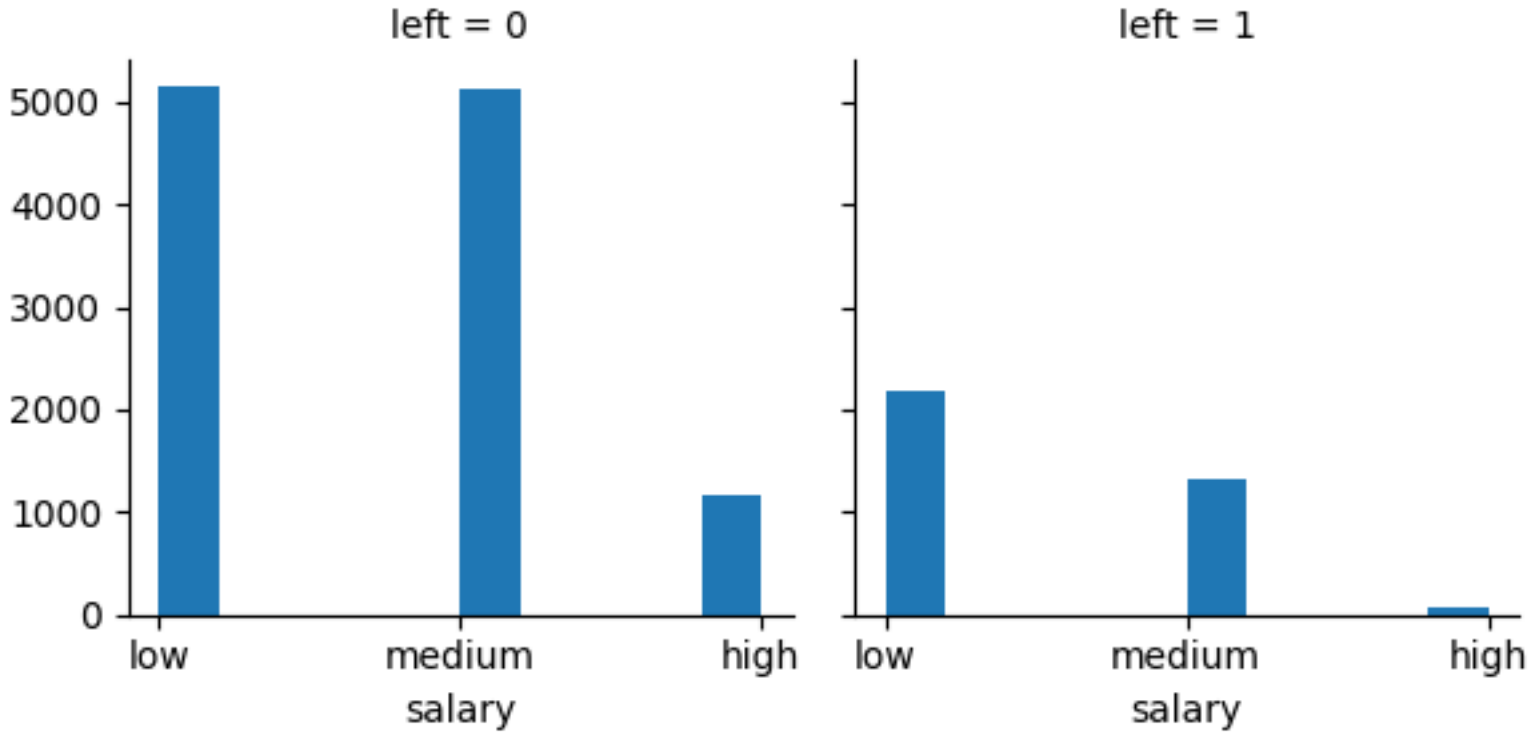

read data
data visualization
correlation analysis
model building
hyperparameter tuning
feature selection
performance
evaluation
No
Yes
feature importance
draw conclusion

Data Correlation



Data Distribution

Salary



	salary	left
1	low	0.296884
2	medium	0.204313
0	high	0.066289

P value = 0
Strong correlation

High salary is more likely to keep employees.

Data Distribution

department

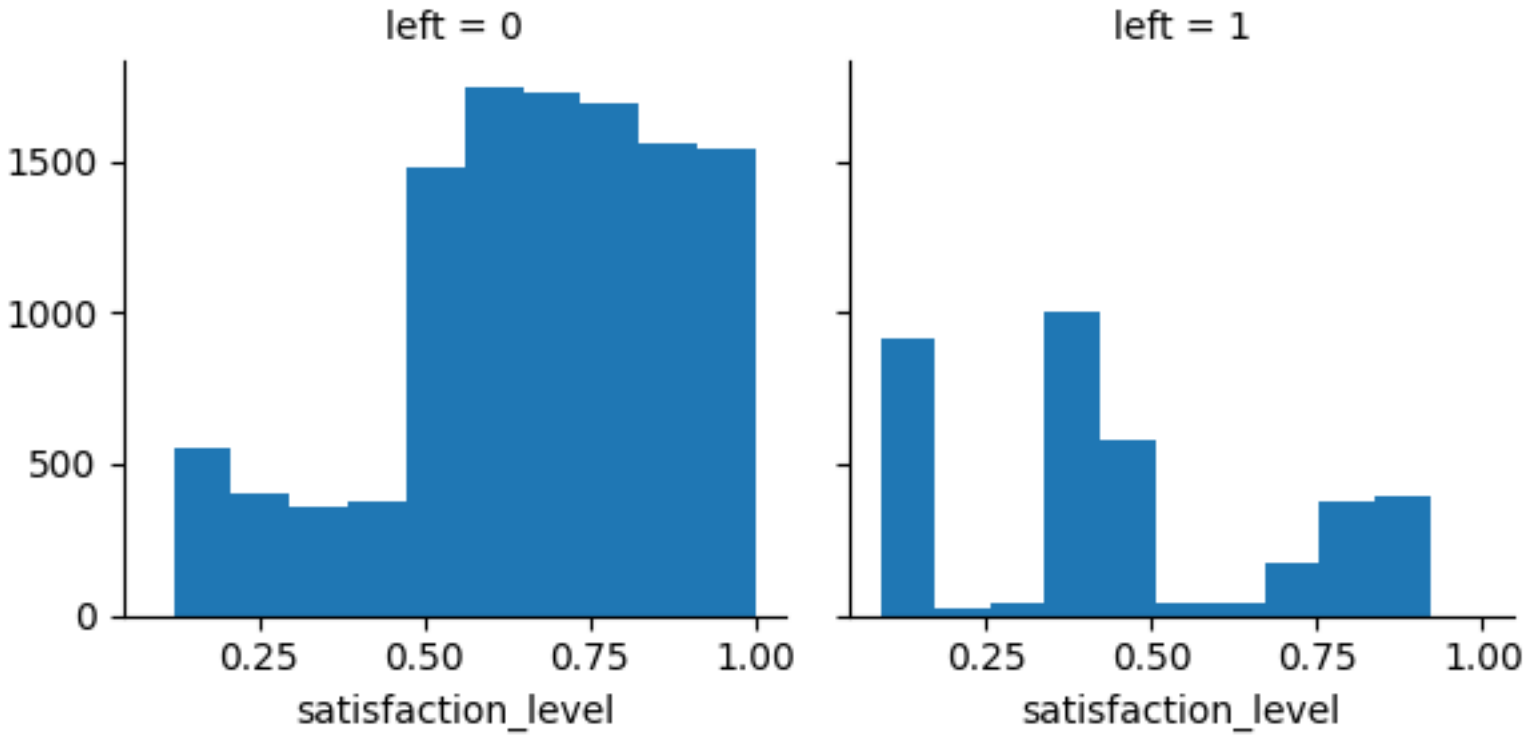
	sales	left
3	hr	0.290934
2	accounting	0.265971
9	technical	0.256250
8	support	0.248991
7	sales	0.244928
5	marketing	0.236597
0	IT	0.222494
6	product_mng	0.219512
1	RandD	0.153748
4	management	0.144444

P value = 0
Strong correlation

Employees in HR, accounting department are the mostly likely to leave.

Data Distribution

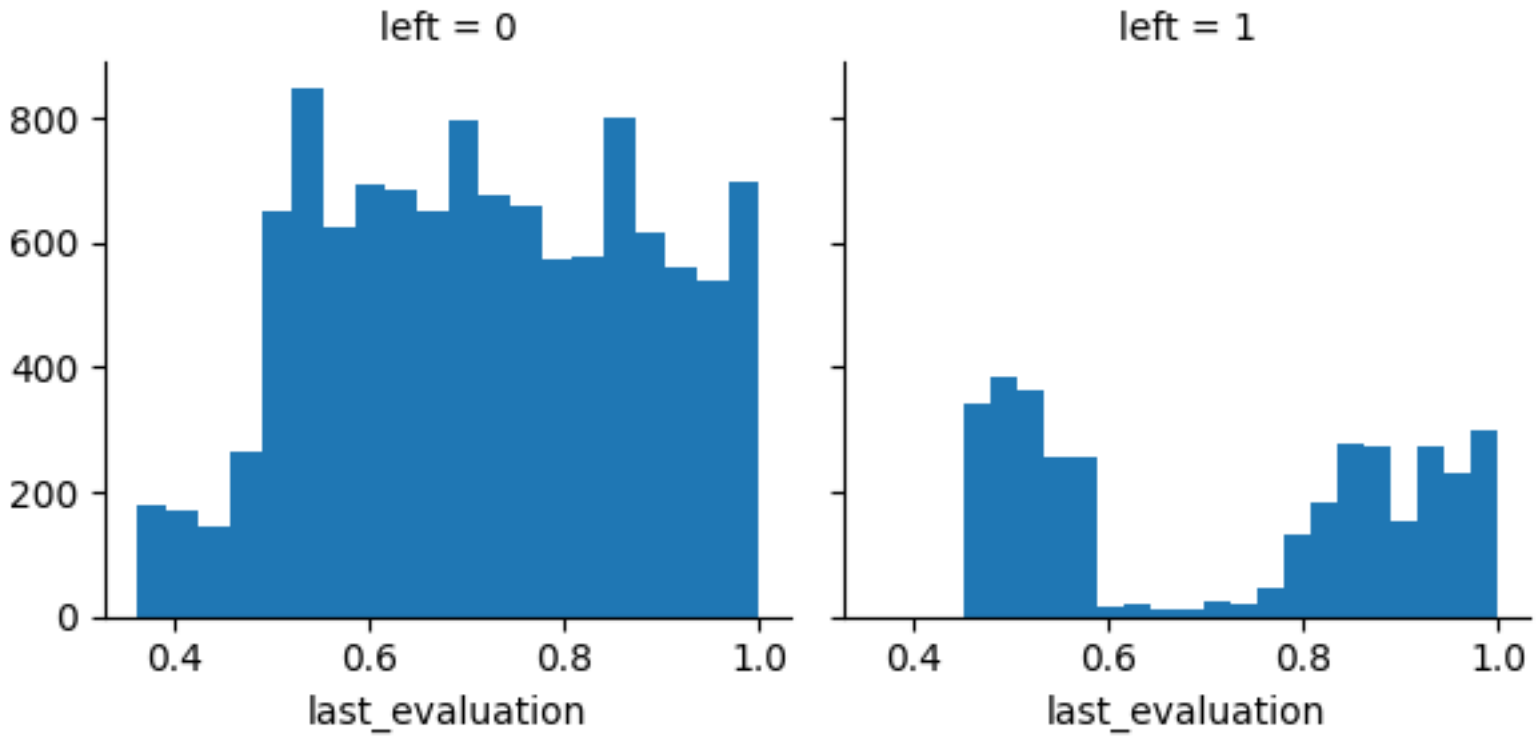
satisfaction_level



P value = 0
Strong correlation

Data Distribution

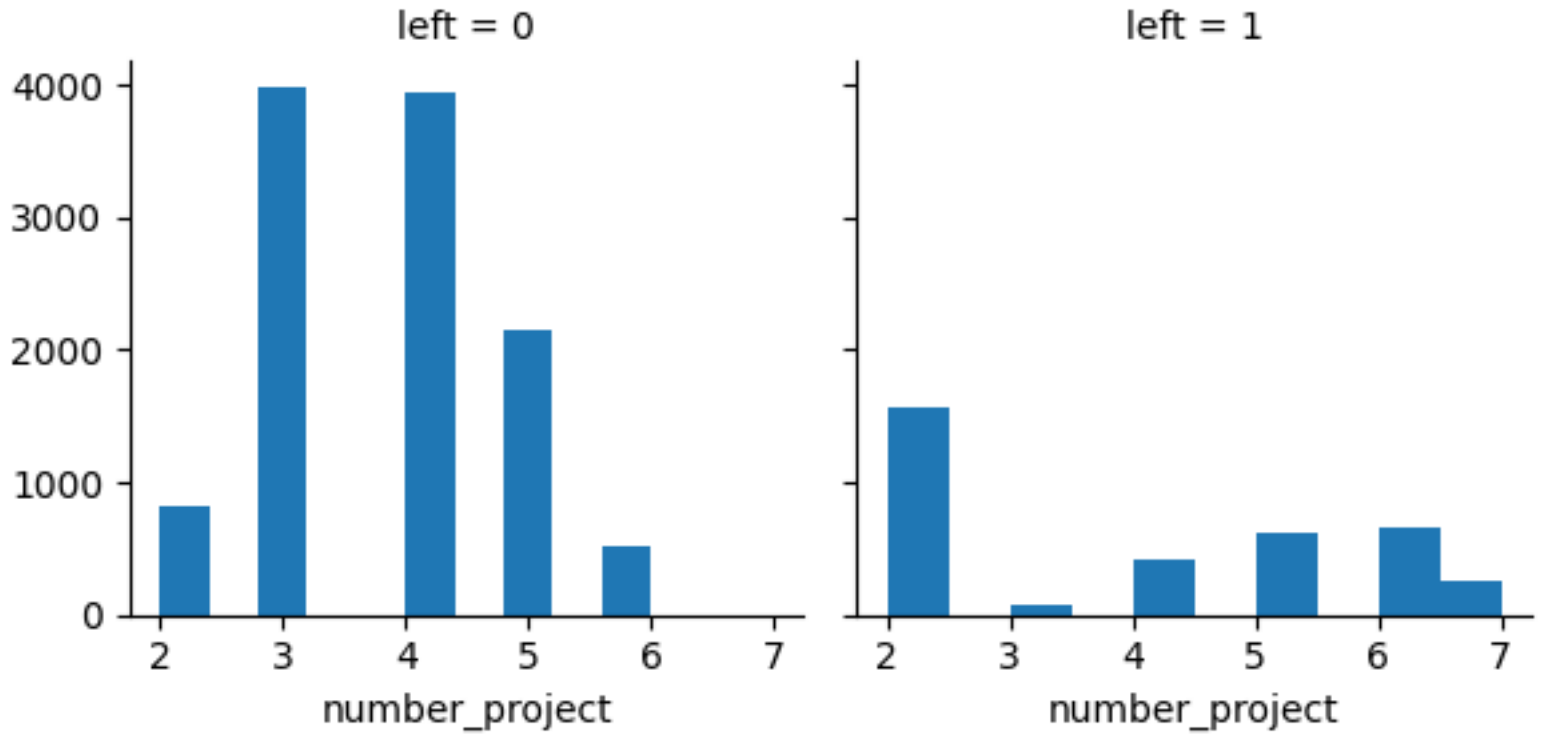
Last_evaluation



P value = 0
Strong correlation

Data Distribution

number_project

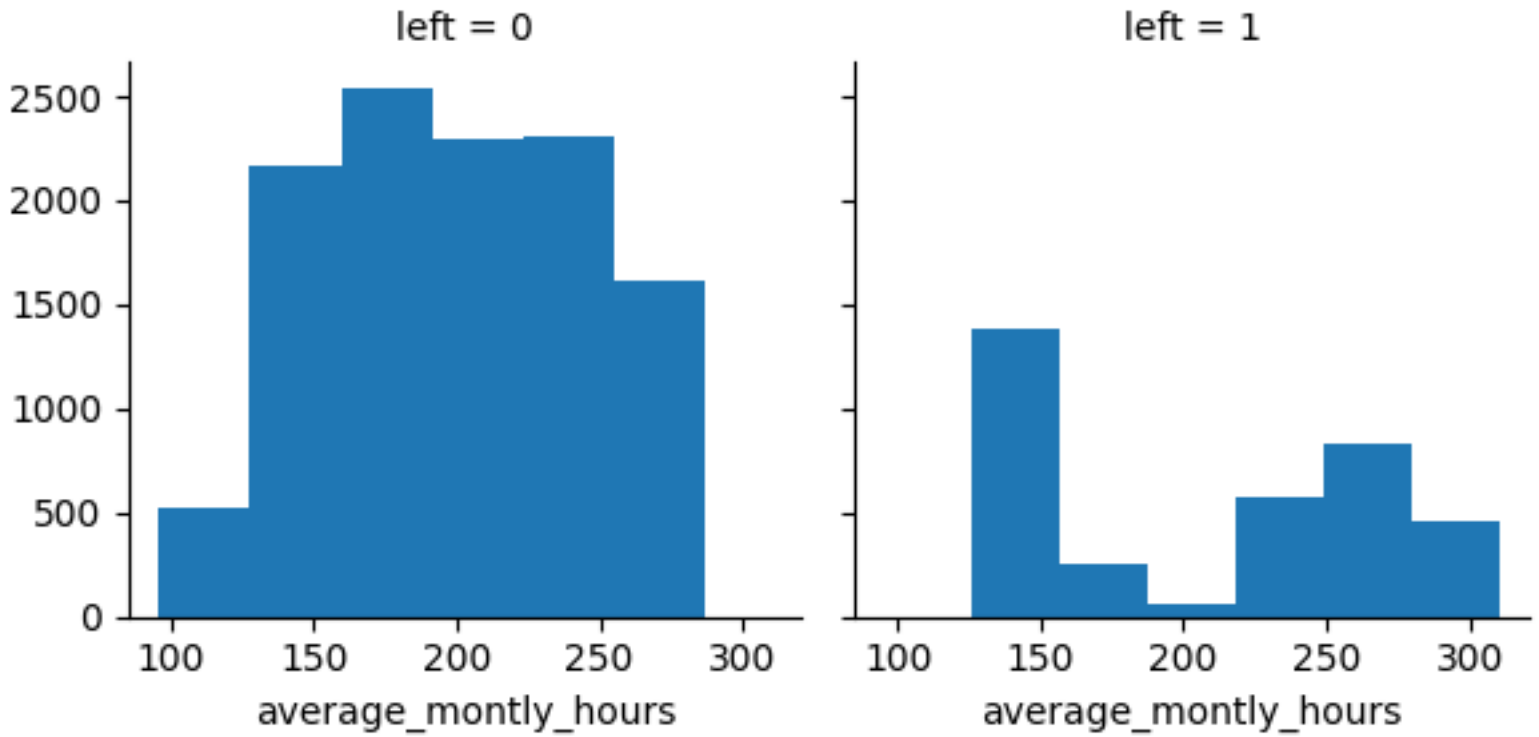


P value = 0
Strong correlation

More number of project might make employees more likely to leave.

Data Distribution

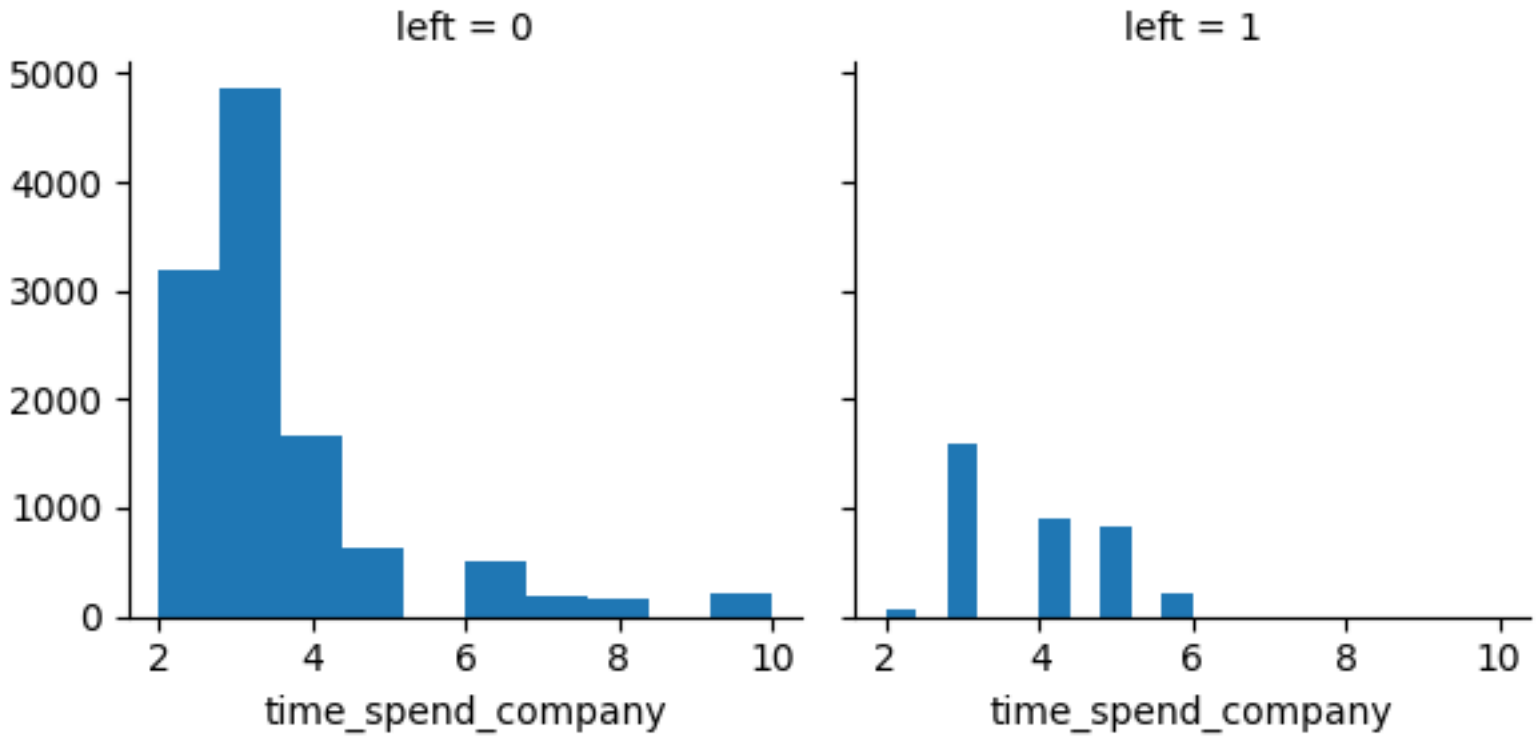
average_montly_hours



P value = 0
Strong correlation

Data Distribution

time_spend_company



P value = 0
Strong correlation

Data Distribution

Other

	Work_accident	left
0	0	0.265160
1	1	0.077916

P value = 0
Strong correlation

	promotion_last_5years	left
0	0	0.241962
1	1	0.059561

P value = 0
Strong correlation

Result

- Baseline

Model Name	Accuracy
XGBoost	0.9876
GBDT	0.9763
Random Forest	0.9904
AdaBoost	0.9602
Extra Trees	0.9860
Decision Tree	0.9782

Result

- Best model
  - Test set recall as the final metric
  - XGBoost
  - Grid Search for hypermeters tuning
  - weight of evidence encoding on ‘department’

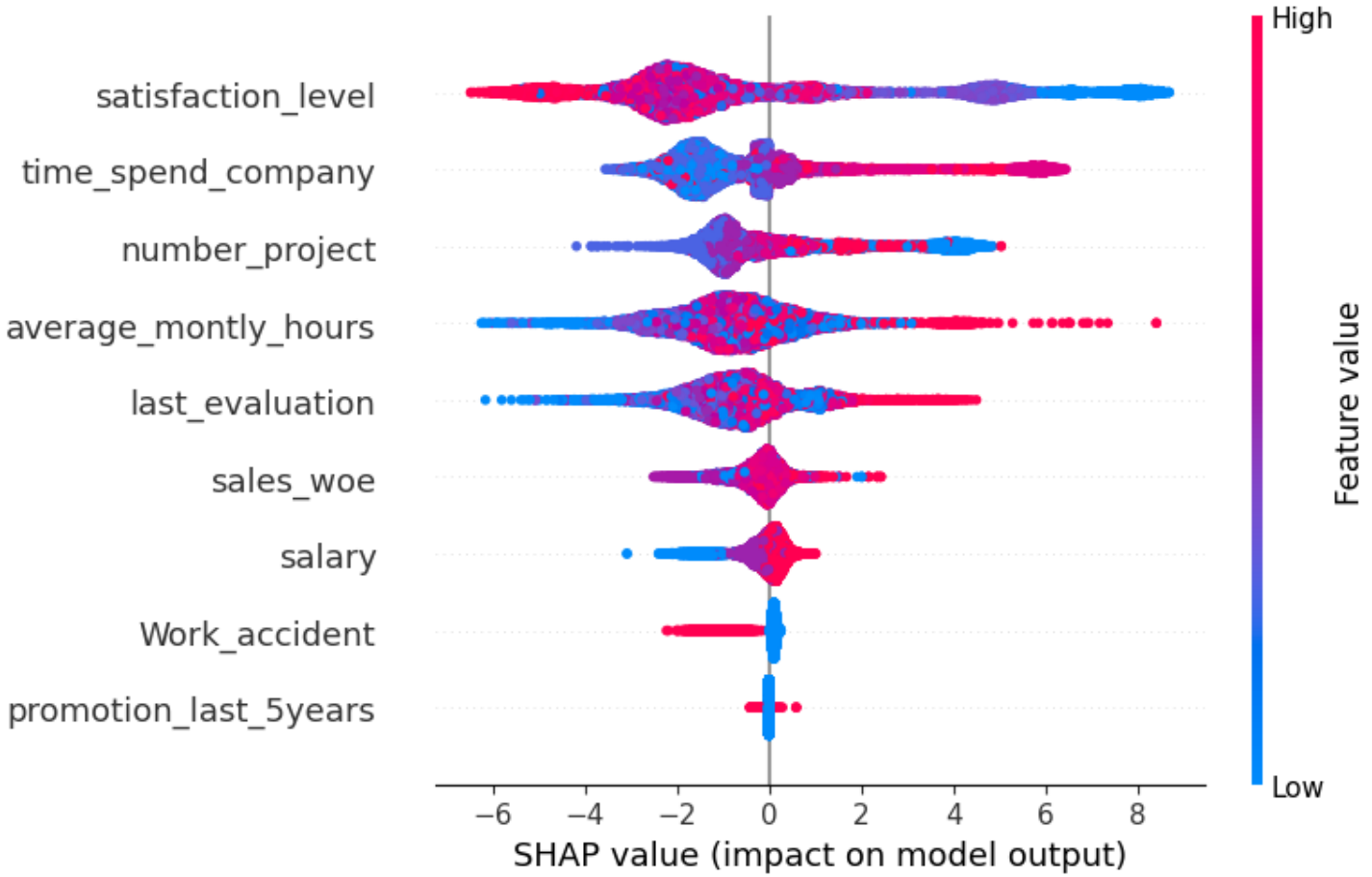
```
Test Set:
Xgboost: 0.9887
Summary
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      2286
           1       0.98      0.97      0.98       714

    accuracy                           0.99      3000
   macro avg       0.99      0.98      0.98      3000
weighted avg       0.99      0.99      0.99      3000

Confusion Matrix:
 [[2274   12]
 [  22  692]]
Test Set recall: 0.96919
```

Feature Importance



- **Satisfaction level** is the most importance feature that influences employees' left
- The lower satisfaction level, the higher left probability.

- Then the **number of project**, **time spent in company**, **average monthly hours** are importance features and are intertwined.
- The more working hours, project number, the higher left probability.

Conclusion

- Understand why employees are leaving the company
  - Longer working hours, more projects will cause lower satisfaction level, and make employees more likely to leave.
- Who will be the next ones to leave
  - Lower salary, employees in HR and accounting department, long working hours, 7 projects in process, these employees might be the next ones to leave.
- Find an action plan to tackle this problem
  - Reduce the project number for each employee, reduce their monthly working time.
  - Raise salary.