HR Analytics

Main purpose of this notebook is to be able to predict employee attrition and fair compensation value.

Contents:

Understanding the HR dataset

Variable	Description
satisfaction_level	Employee-reported job satisfaction level [0–1]
last_evaluation	Score of employee's last performance review [0–1]
number_project	Number of projects employee contributes to
average_monthly_hours	Average number of hours employee worked per month
time_spend_company	How long the employee has been with the company (years)
Work_accident	Whether or not the employee experienced an accident while at work
left	Whether or not the employee left the company
promotion_last_5years	Whether or not the employee was promoted in the last 5 years
Department	The employee's department
salary	The employee's salary (U.S. dollars)

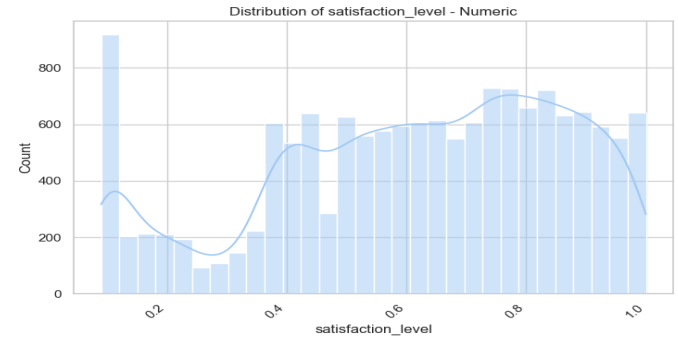
Initial EDA and data cleaning

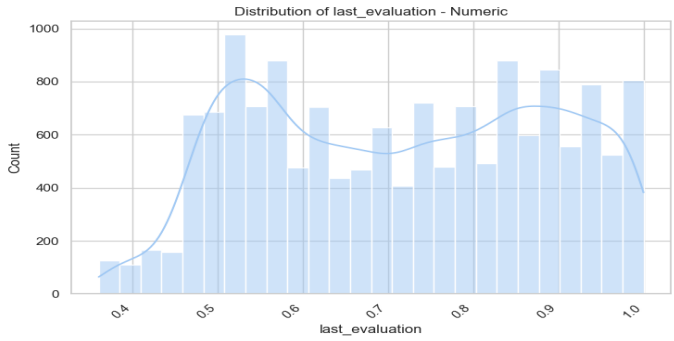
Data Information post data type change:

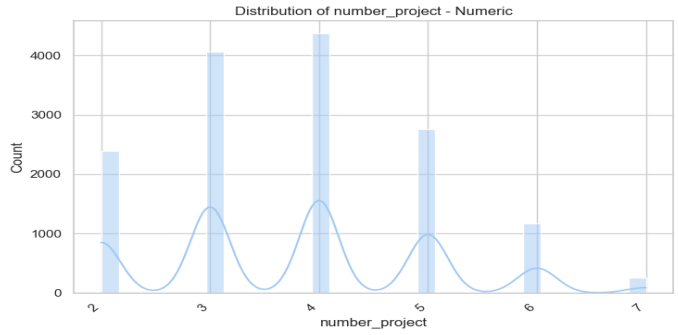
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14999 entries, 0 to 14998
Data columns (total 10 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   satisfaction_level     14999 non-null  float64
 1   last_evaluation        14999 non-null  float64
 2   number_project         14999 non-null  int64
 3   average_monthly_hours  14999 non-null  int64
 4   tenure                 14999 non-null  int64
 5   work_accident          14999 non-null  int64
 6   left                   14999 non-null  int64
 7   promotion_last_5years  14999 non-null  int64
 8   department             14999 non-null  object
 9   salary_usd             14999 non-null  int64
dtypes: float64(2), int64(7), object(1)
memory usage: 1.1+ MB
```

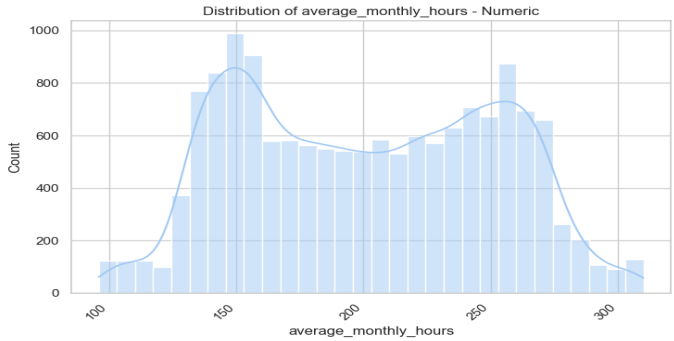
Number of missing rows per column

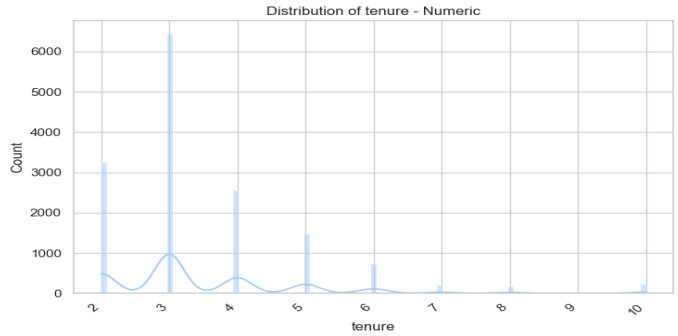
```
satisfaction_level       0
last_evaluation          0
number_project           0
average_monthly_hours    0
tenure                   0
work_accident            0
left                     0
promotion_last_5years    0
department               0
salary_usd               0
dtype: int64
```

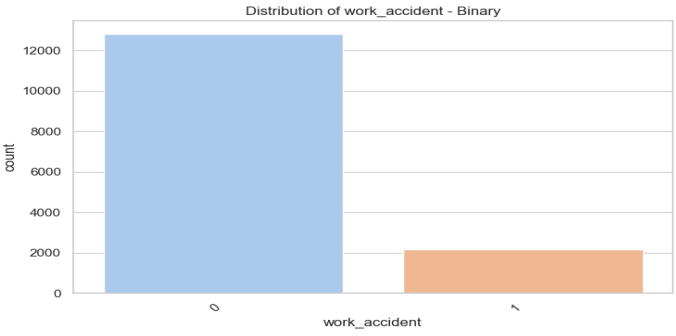

Distribution of satisfaction_level - Numeric
Count
satisfaction_level

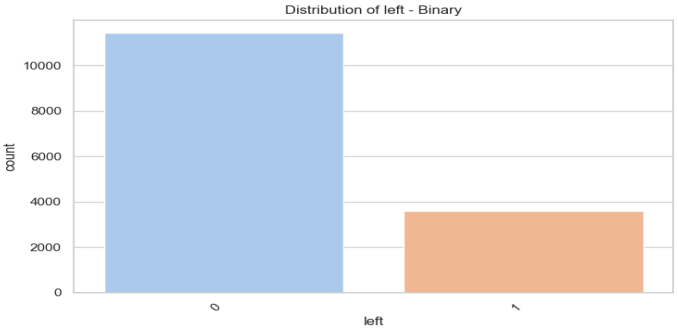

Distribution of last_evaluation - Numeric
Count
last_evaluation

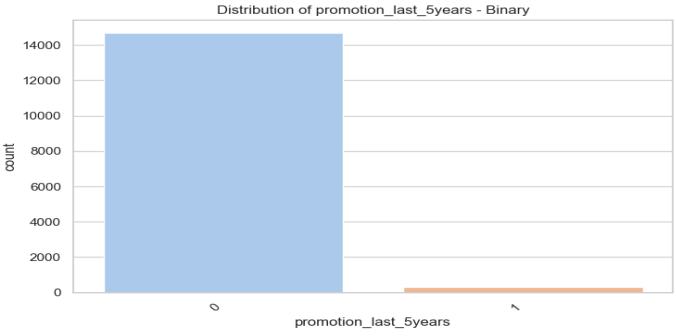

Distribution of number_project - Numeric
Count
number_project

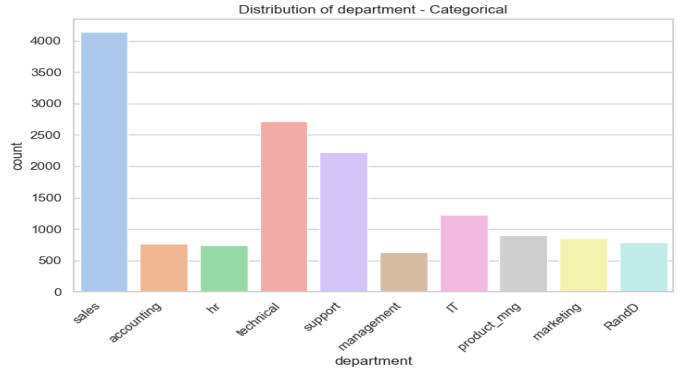

Distribution of average_monthly_hours - Numeric
Count
average_monthly_hours

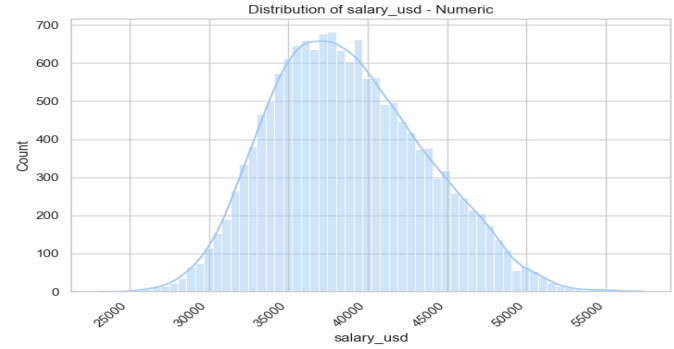

Distribution of tenure - Numeric
Count
tenure


Distribution of work_accident - Binary
count
work_accident


Distribution of left - Binary
count
left


Distribution of promotion_last_5years - Binary
count
promotion_last_5years


Distribution of department - Categorical
count
sales
accounting
hr
technical
support
management
IT
product_mng
marketing
RandD
department


Distribution of salary_usd - Numeric
Count
salary_usd

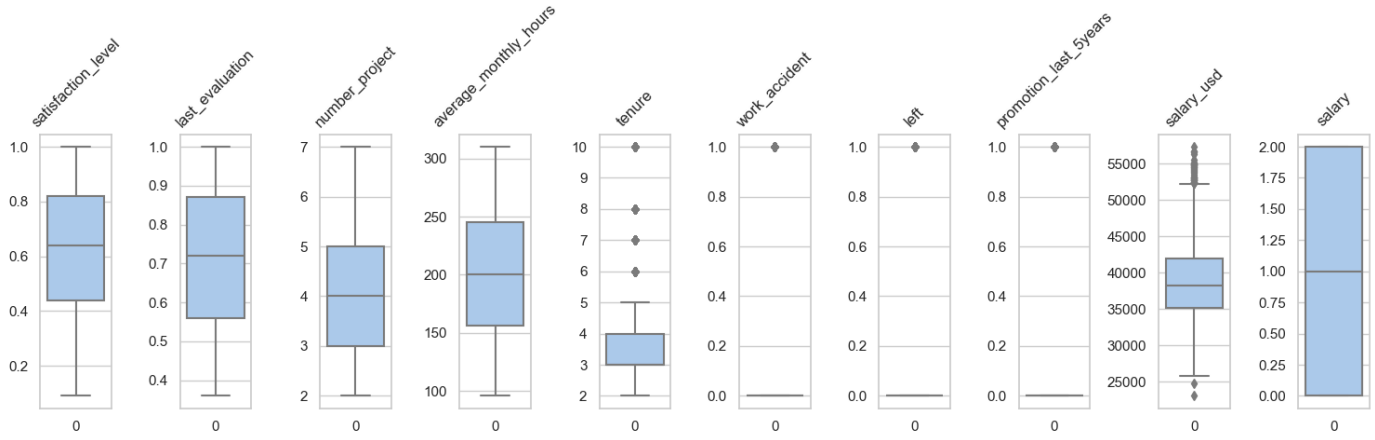
Initial observations of the data:

* About 14% of employees have experienced a work accident

* Almost 98% of employees did not receive a promotion in the last 5 years

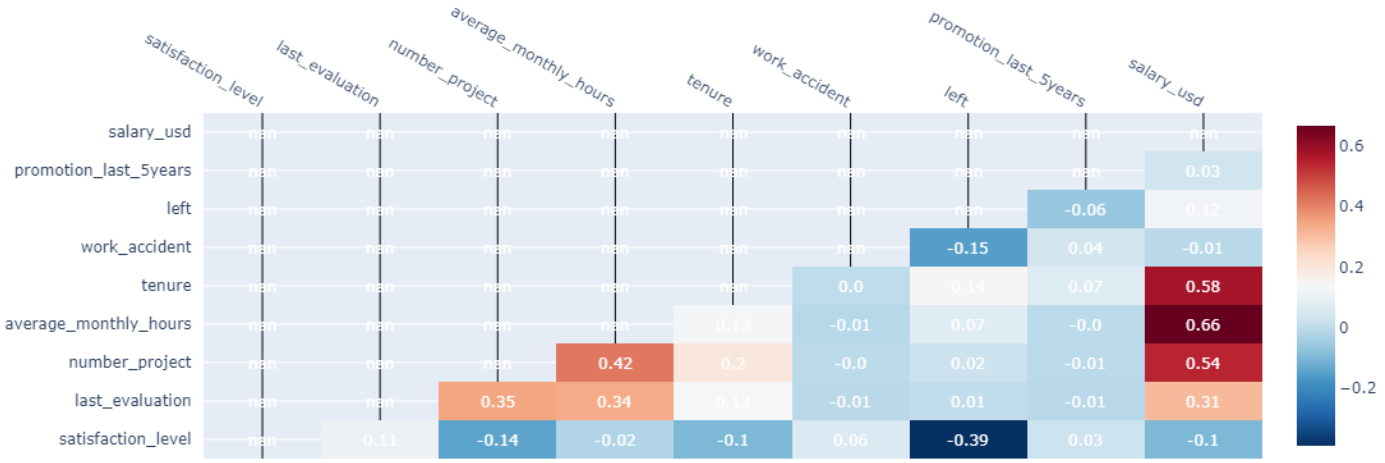
* Almost 24% of employees have left the company

Outliers



	Number of Outliers	Percentage
left	3571	23.808254
work_accident	2169	14.460964
tenure	1282	8.547236
promotion_last_5years	319	2.126808
salary_usd	46	0.306687
satisfaction_level	0	0.000000
last_evaluation	0	0.000000
number_project	0	0.000000
average_monthly_hours	0	0.000000
salary	0	0.000000

Correlation analysis



There are some important correlations to be noted:

1. salary_usd with 66% positive correlation with average monthly hours, 58% with tenure and 54% with number_project

2. left and satisfaction_level: -39% negative correlation

3. average_monthy_hours and number_project : 42% correlated

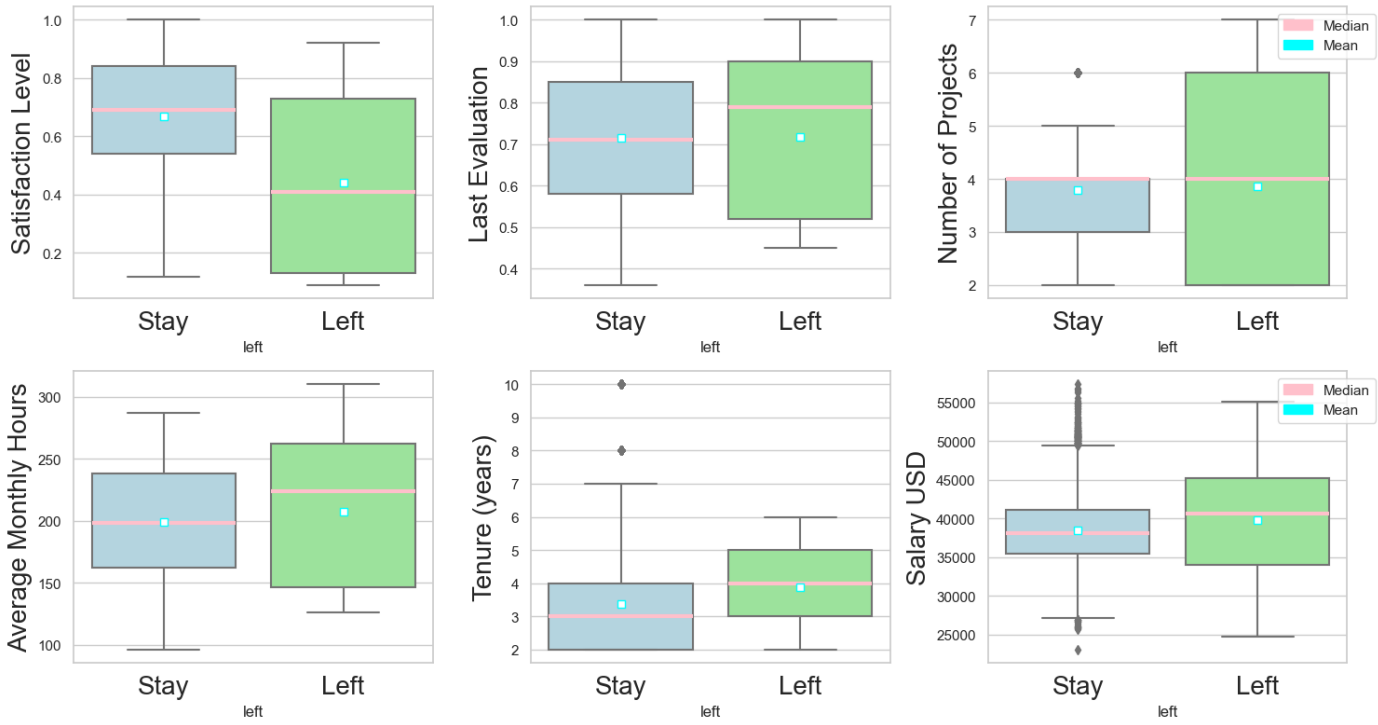
4. number_project and last_evaluation: 35% correlated

5. average_monthly_hours and last_evaluation: 34% correlated

- Tenure: The correlation of 0.15 indicates that there is a weak tendancy for people that have been with the company for a long time to leave.

- Satisfaction level: The negative correlation of -0.38 indicates that less satisfied employees are more likely to leave. However, this is not a very strong inverse correlation.

Analyzing the differences between those who left vs those who stayed



Comparing the characteristics of employees who left with those who stayed. Here are the key findings:

	p-value	statistically significant (a = 0.05)
Numerical features		
satisfaction_level	0.000000e+00	True
tenure	4.207680e-71	True
salary_usd	1.000342e-47	True
average_monthly_hours	2.311304e-18	True
number_project	3.575214e-03	True
last_evaluation	4.212702e-01	False

- **Satisfaction Level:** There is a statistically significant difference in the asatisfaction levels of employees who left and those who stayed. Employees who left the company tend to have a lower satisfaction level.

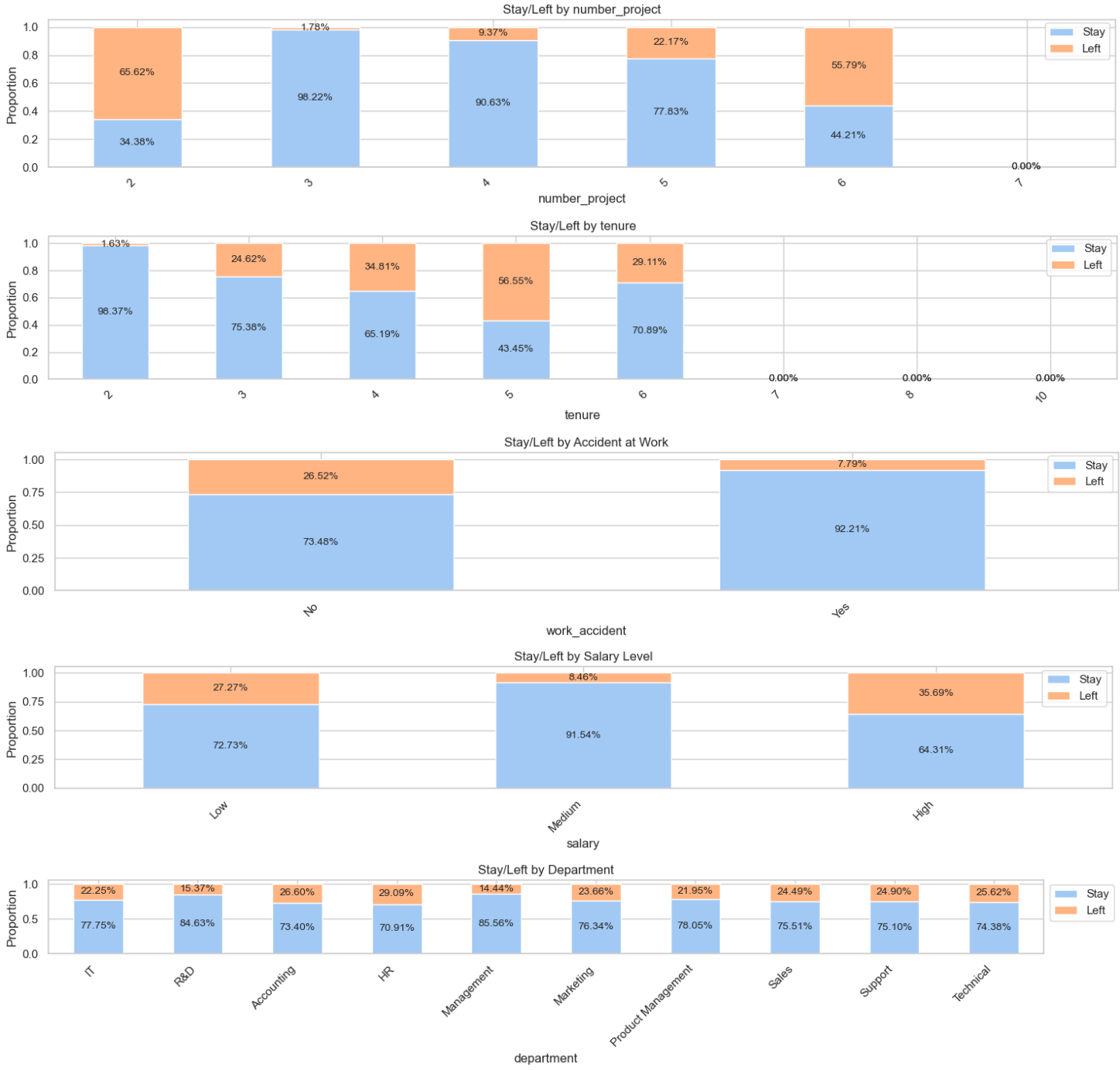
- **Tenure (years):** There's a statistically significant difference in tenure between the groups. Employees who left have slightly higher tenure compared to those who stayed.

- **Salary Level:** There's a significant difference in the salary levels of the two groups.

- **Average Monthly Hours**: Employees who left tend to work more hours on average, and this difference is statistically significant.

- **Number of Projects**: Employees who left have a statistically significant difference in the number of projects they were involved in.

- **Last Evaluation:** While there's a slight difference in the last evaluation scores between the two groups, this difference is not statistically significant. This suggests that the quality of work (as measured by the last evaluation) might not be a strong predictor for an employee's decision to leave.



p-value: 0.0 : The difference in proportions between number_project and left is statistically significant.

p-value: 0.0 : The difference in proportions between tenure and left is statistically significant.

p-value: 9.55823958002199e-80 : The difference in proportions between work_accident and left is statistically significant.

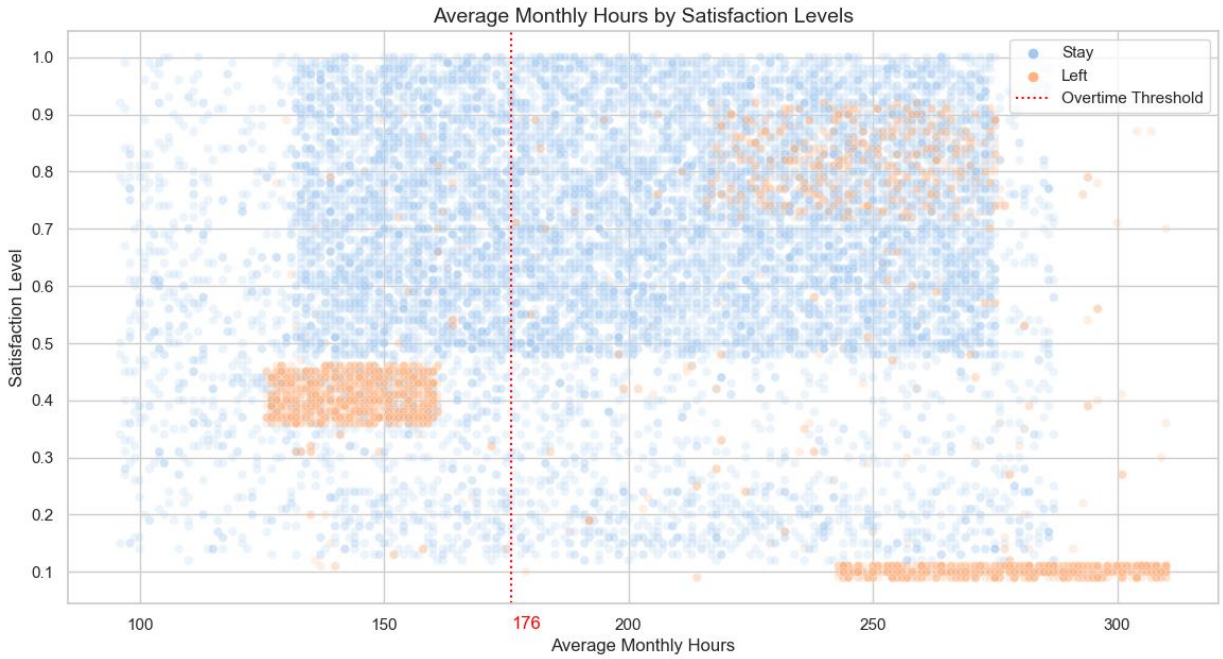
p-value: 2.3223449753167188e-233 : The difference in proportions between salary and left is statistically significant.

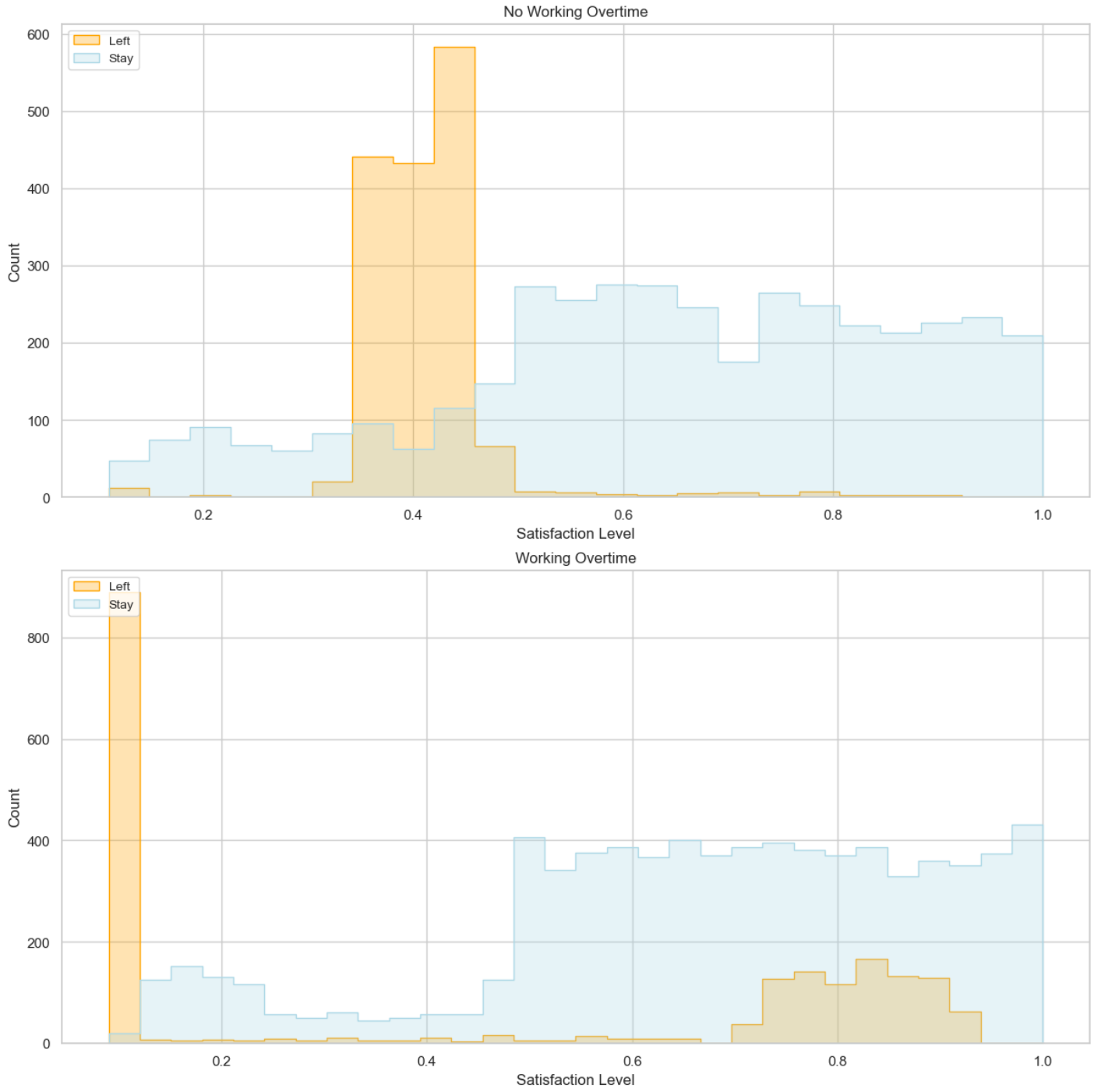
p-value: 7.042130463822518e-15 : The difference in proportions between department and left is statistically significant.

Findings

1. All the variables except last_evaluation are statistically significant between employees who left vs those that stayed.
2. Surprisingly, the proportion of employees leaving is lower for those who had an accident at work (5.68%) compared to those who did not (18.60%).
3. The proportion of employees leaving is lower for those who received a promotion in the last five years i.e. 3.94% compared to those who did not receive a promotion i.e. 16.82%.
4. The proportion of employees leaving is the lowest among the high salary group i.e. 4.85%, followed by the medium salary group i.e. 14.62% and the low salary group i.e. 20.45%. This relationship is also depicted in the previous boxplot visualization of salary.

Overtime analysis





From the histograms, we can observe the distribution of satisfaction levels among employees based on whether they worked overtime and whether they stayed with the company or left.

- Employees not working overtime tend to leave when their satisfaction level is around 0.4.

- Among those doing overtime, two groups are prominent: one with very low satisfaction and another around a 0.8 satisfaction level, both showing a higher likelihood to quit.

Modelling

Linear Regression for predicting salary

Assumptions

- **Linearity**: The relationship between the predictors and the target variable should be linear. We'll use scatter plots for some of the continuous variables against the 'salary_usd' to visually inspect linearity.

- **Homoscedasticity**: The residuals (differences between observed and predicted values) should have constant variance.

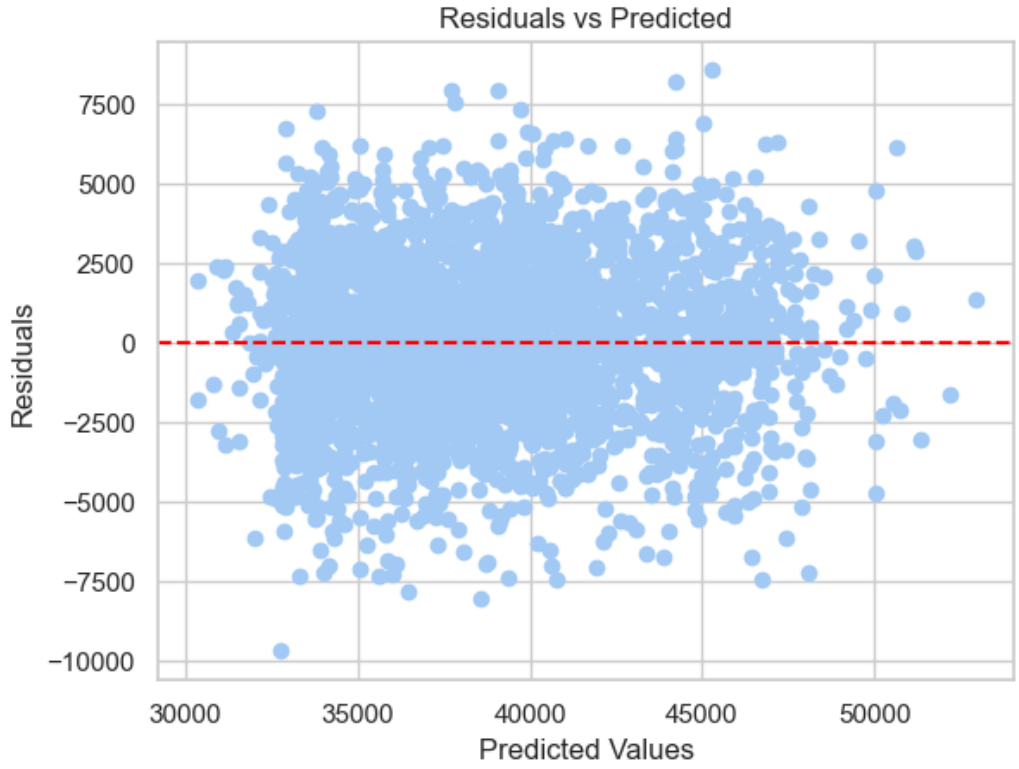
- **Normality** of Residuals: The residuals should be normally distributed.

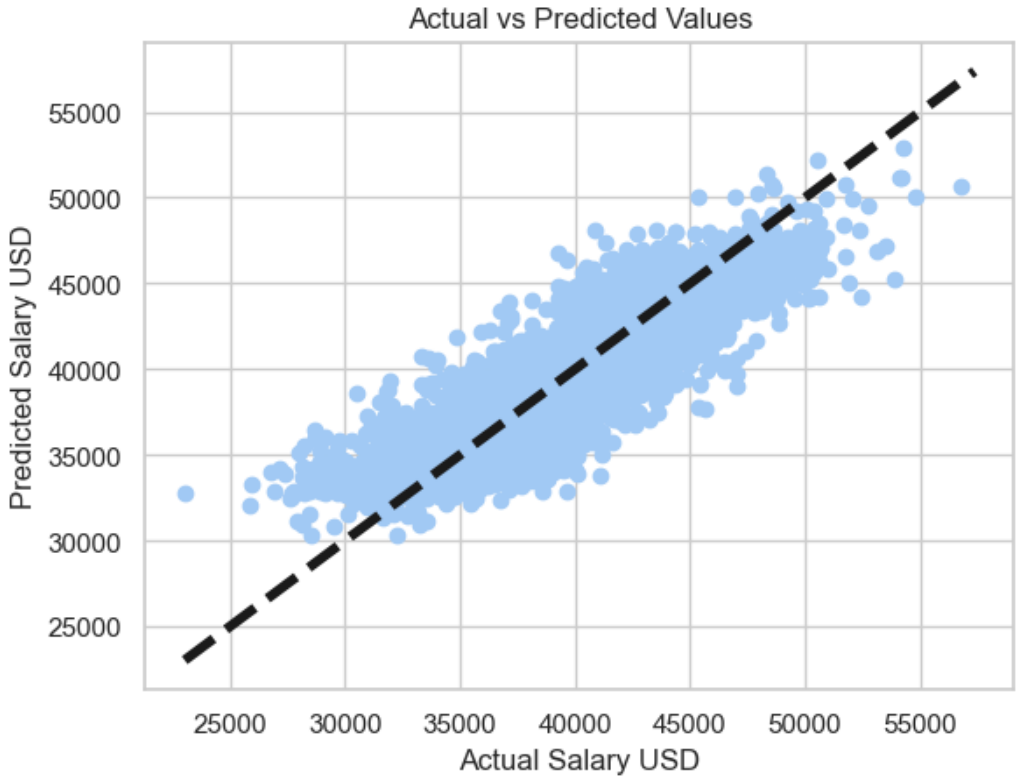
- **Independence** of Residuals: Residuals should be independent of each other.

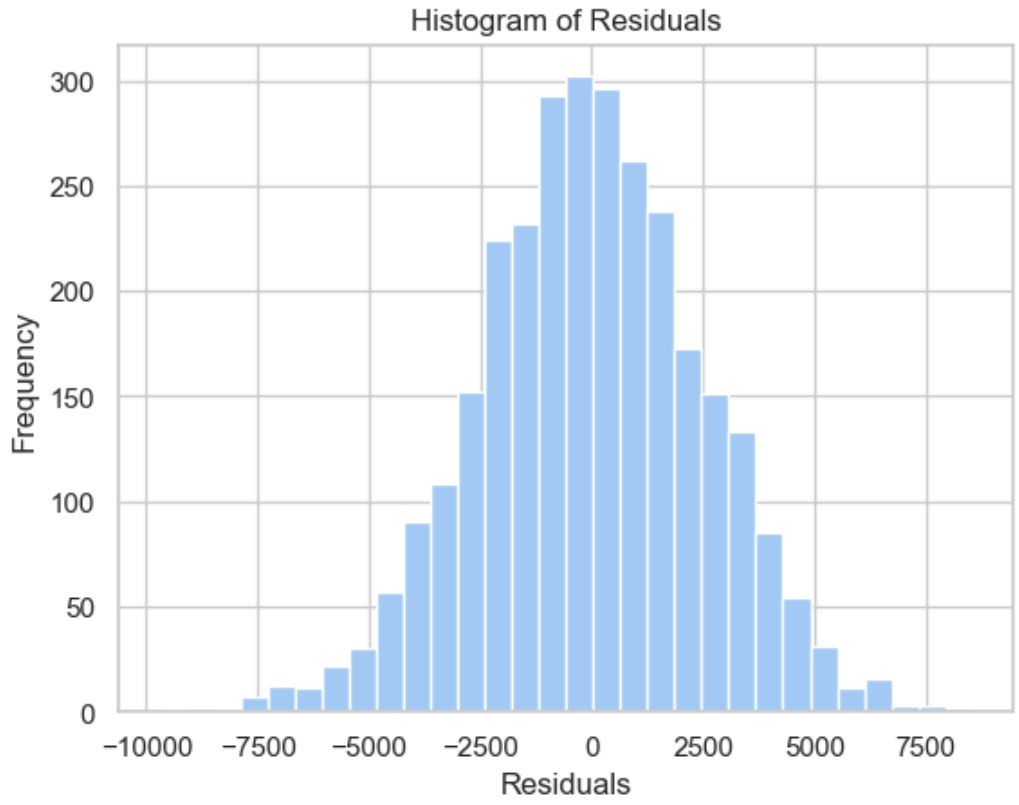
```
Selected features: ['number_project', 'average_monthly_hours', 'tenure', 'left', 'overtime']
                            OLS Regression Results
==============================================================================
Dep. Variable:             salary_usd   R-squared:                       0.872
Model:                            OLS   Adj. R-squared:                  0.872
Method:                 Least Squares   F-statistic:                 1.699e+04
Date:                Sun, 31 Dec 2023   Prob (F-statistic):               0.00
Time:                        16:14:17   Log-Likelihood:            -1.3316e+05
No. Observations:               14999   AIC:                         2.663e+05
Df Residuals:                   14992   BIC:                         2.664e+05
Df Model:                           6
Covariance Type:            nonrobust
===========================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------
const                    2.654e+04    100.544    264.014      0.000    2.63e+04    2.67e+04
number_project            436.8108     13.429     32.527      0.000     410.488     463.133
average_monthly_hours      21.7869      0.592     36.833      0.000      20.627      22.946
tenure                    764.5177     11.728     65.188      0.000     741.530     787.506
left                      180.7226     35.085      5.151      0.000     111.952     249.493
salary                   3554.3557     27.949    127.175      0.000    3499.573    3609.138
overtime                 -206.7494     55.239     -3.743      0.000    -315.025     -98.473
==============================================================================
Omnibus:                      269.372   Durbin-Watson:                   1.999
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              547.434
Skew:                           0.049   Prob(JB):                    1.34e-119
Kurtosis:                       3.931   Cond. No.                     1.56e+03
==============================================================================
```

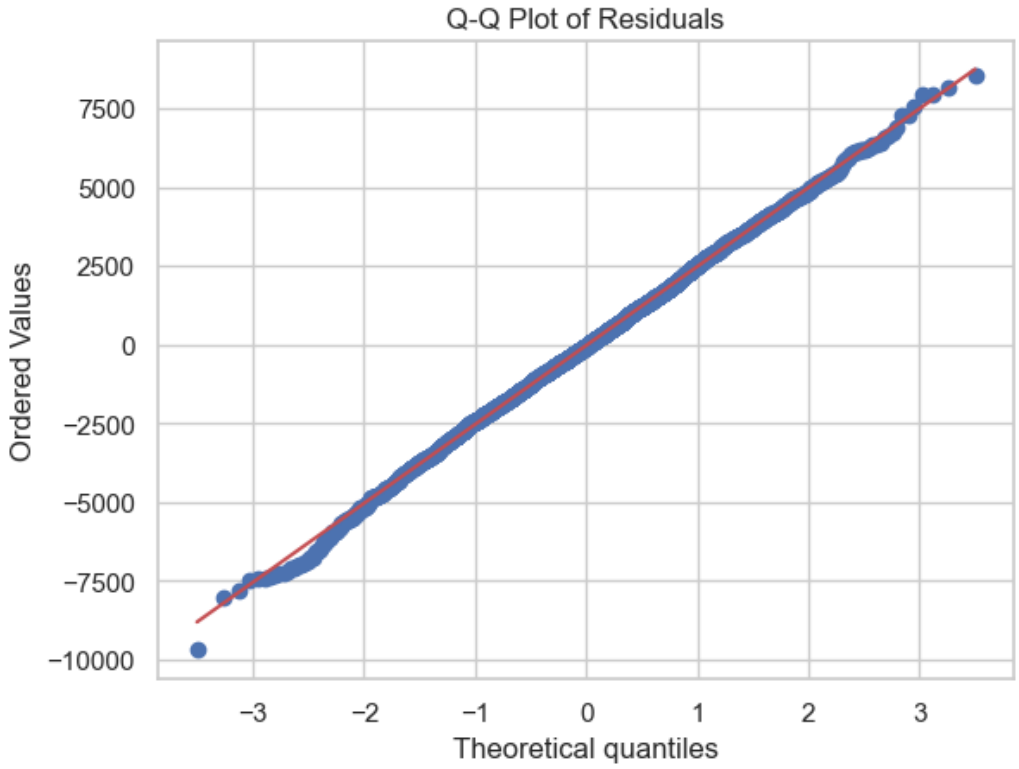
After splitting the data into training and test sets, the results are the following:

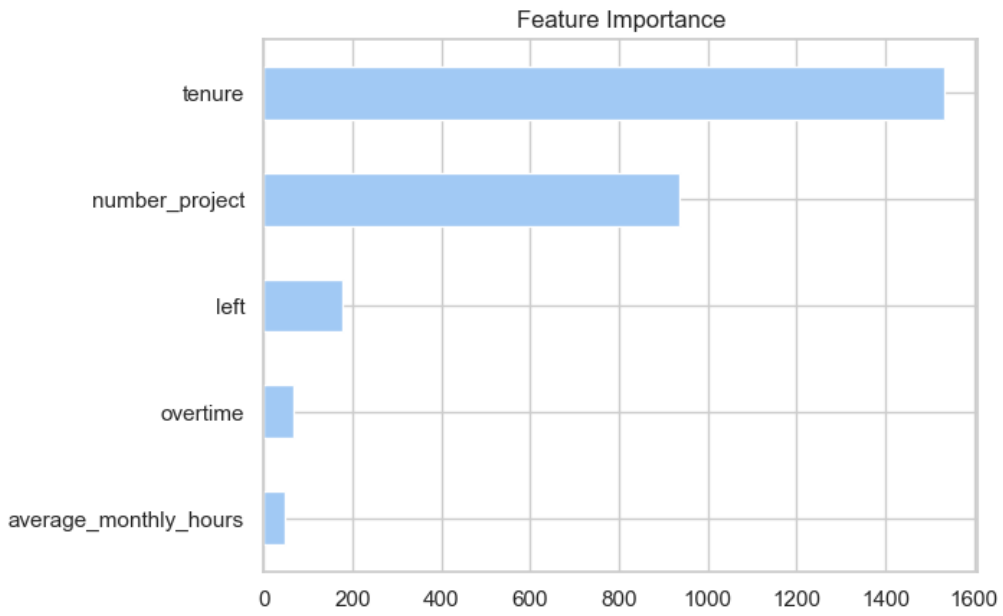
R-squared: 0.7390454428056037, RMSE: 2504.807841306232






Histogram of Residuals
Frequency
Residuals
300
250
200
150
100
50
0
−10000
−7500
−5000
−2500
0
2500
5000
7500


Q-Q Plot of Residuals
Ordered Values
Theoretical quantiles
7500
5000
2500
0
−2500
−5000
−7500
−10000
−3
−2
−1
0
1
2
3



Linear Regression Conclusions:

- **Model Performance**: The R-squared value of approximately 0.739 suggests that the model explains about 73.9% of the variance in the salary data, which is a strong level of explanation given the complexity inherent in salary determinations.

- **Residual Analysis**: The "Residuals vs. Predicted" plot indicates a reasonable spread around the zero line, suggesting that the model's predictions are unbiased on average. However, the slight pattern of increasing spread with higher predicted values hints at heteroscedasticity, implying that the model's precision decreases as salary increases.

- **Predictive Accuracy**: The "Actual vs. Predicted Values" plot reveals that the model's predictions are generally close to the actual salaries, especially in the middle range of salaries. However, there is visible deviation as the actual salary increases, which may signal that the model is less accurate for higher salary ranges.

- **Distribution of Residuals**: The histogram and Q-Q plot of residuals display a roughly normal distribution, with some minor deviations in the tails. This slight skewness and the tail behavior observed in the Q-Q plot suggest that extreme values are not as well predicted by the model, potentially affecting the accuracy of predictions and the width of the confidence intervals.

- **Feature Importance**: The "Feature Importance" plot identifies 'tenure' and 'number_project' as the most significant predictors of salary. This is consistent with domain understanding, as tenure can be associated with increased experience and career progression, which typically results in higher salaries. The number of projects may reflect an employee's level of engagement and responsibility, which also correlates with compensation.

- **RMSE**: The RMSE of 2504.81 indicates that the typical prediction error is approximately $2,505. While this is a relatively small error in the context of salary predictions, it is still significant and suggests room for improvement in the model's predictive power.

- **Confidence Intervals**: The wide confidence intervals compared to the actual values, particularly for some predictions, suggest that certain predictions made by the model are associated with substantial uncertainty. This could be addressed by exploring model improvements or alternative modeling approaches.

Predicting Attrition

Logistic Regression

```
The outcome variable is binary or categorical.
Ensure that the observations are independent.
                   feature         VIF
0       satisfaction_level    7.463931
1          last_evaluation   22.638355
2           number_project   16.351013
3    average_monthly_hours   79.110191
4                   tenure   11.985591
5            work_accident    1.177653
6    promotion_last_5years    1.052055
7              salary_usd  128.468777
8                  salary    5.584996
9                overtime    8.809346
10       department_RandD    1.620013
11   department_accounting    1.605254
12          department_hr    1.574656
13   department_management    1.536524
14    department_marketing    1.684781
15  department_product_mng    1.709421
16        department_sales    4.259137
17      department_support    2.756707
18    department_technical    3.133659
Found 674 extreme outliers
The sample size is sufficiently large.
```

Only the features with a VIF < 10 were chosen for logistic regression.

	No Scaling				MinMaxScaler				RobustScaler				StandardScaler			
	precision	recall	f1-score	support	precision	recall	f1-score	support	precision	recall	f1-score	support	precision	recall	f1-score	support
Stay	0.881808	0.708224	0.785541	2286.000000	0.885730	0.708661	0.787363	2286.000000	0.894626	0.742782	0.811663	2286.000000	0.886653	0.732283	0.802108	2286.000000
Left	0.426976	0.696078	0.529286	714.000000	0.431255	0.707283	0.535809	714.000000	0.466425	0.719888	0.566079	714.000000	0.449640	0.700280	0.547645	714.000000
accuracy	0.705333	0.705333	0.705333	0.705333	0.708333	0.708333	0.708333	0.708333	0.737333	0.737333	0.737333	0.737333	0.724667	0.724667	0.724667	0.724667
macro avg	0.654392	0.702151	0.657414	3000.000000	0.658493	0.707972	0.661586	3000.000000	0.680525	0.731335	0.688871	3000.000000	0.668146	0.716282	0.674877	3000.000000
weighted avg	0.773558	0.705333	0.724552	3000.000000	0.777565	0.708333	0.727493	3000.000000	0.792714	0.737333	0.753214	3000.000000	0.782644	0.724667	0.741546	3000.000000

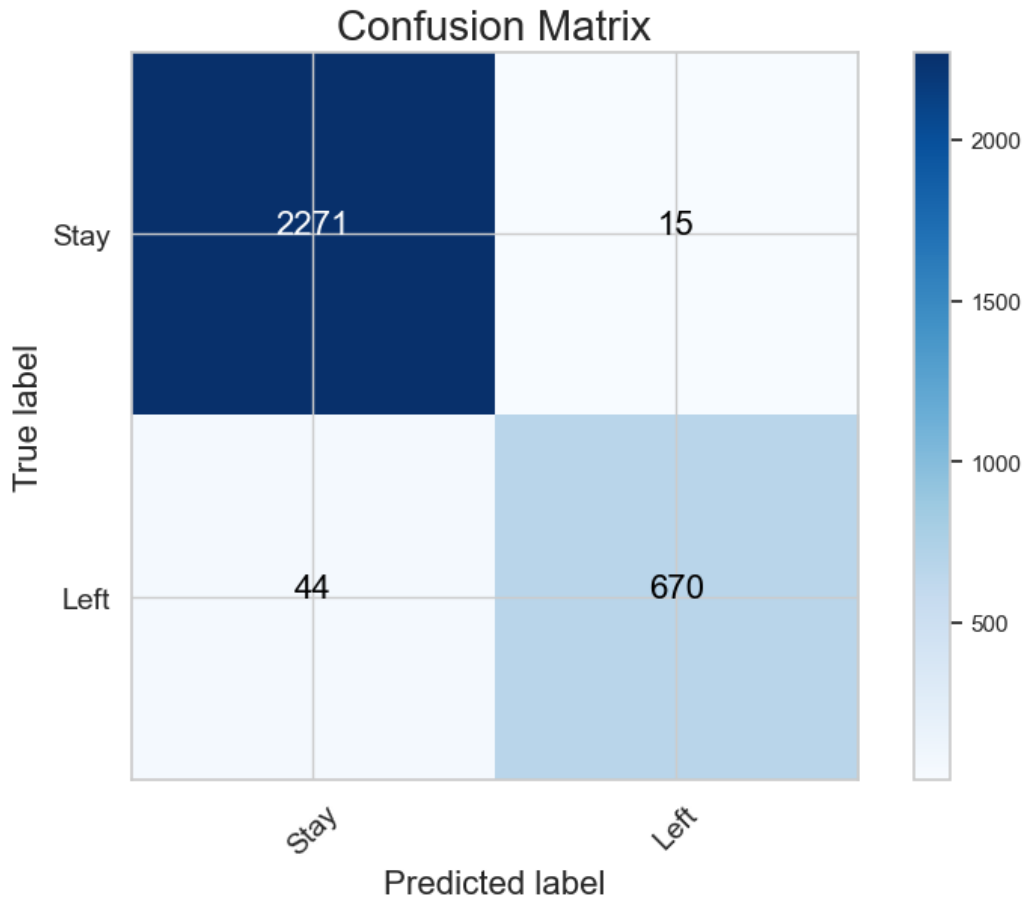
	precision	recall	f1-score	support	StandardScaler
Stay	0.886653	0.732283	0.802108	2286.000000	NaN
Left	0.449640	0.700280	0.547645	714.000000	NaN
accuracy	0.724667	0.724667	0.724667	0.724667	NaN
macro avg	0.668146	0.716282	0.674877	3000.000000	NaN
weighted avg	0.782644	0.724667	0.741546	3000.000000	NaN
ROC AUC	NaN	NaN	NaN	NaN	0.781394
AP	NaN	NaN	NaN	NaN	0.563892
Balanced accuracy	NaN	NaN	NaN	NaN	0.716282
G-mean	NaN	NaN	NaN	NaN	0.716103
Youden's index	NaN	NaN	NaN	NaN	0.432564
MCC	NaN	NaN	NaN	NaN	0.381403
Training_time	NaN	NaN	NaN	NaN	0.195925

- The scaling of data appears to have a marginal impact on the model's performance, with MinMaxScaler and StandardScaler showing slightly better performance metrics compared to no scaling.

- The training time is notably improved with scaled data, which can be a crucial factor in larger datasets or real-time systems.

Tree based clasifiers

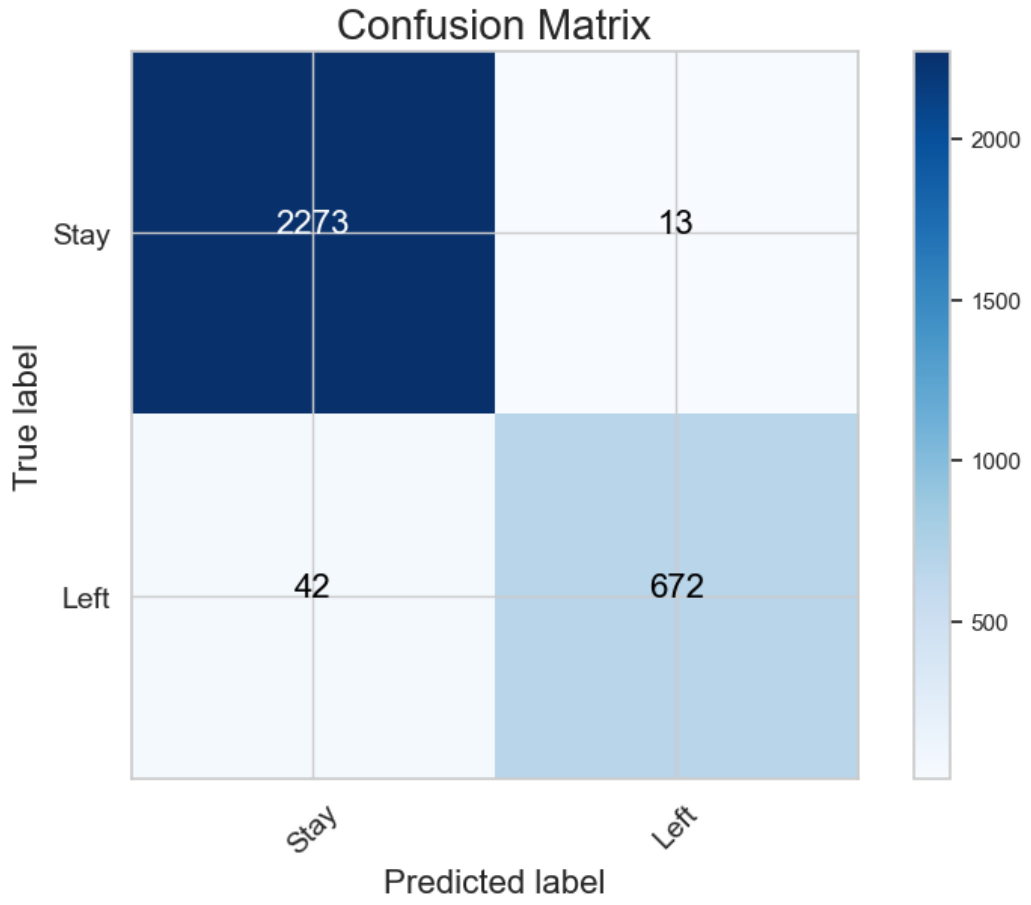
XGBoost



Classification Report for XGBoost

	precision	recall	f1-score	support	ROC AUC	AP	Balanced accuracy	G-mean	Youden's index	MCC
Stay	0.980994	0.993438	0.987177	2286.000000	NaN	NaN	NaN	NaN	NaN	NaN
Left	0.978102	0.938375	0.957827	714.000000	NaN	NaN	NaN	NaN	NaN	NaN
accuracy	0.980333	0.980333	0.980333	0.980333	NaN	NaN	NaN	NaN	NaN	NaN
macro avg	0.979548	0.965907	0.972502	3000.000000	NaN	NaN	NaN	NaN	NaN	NaN
weighted avg	0.980305	0.980333	0.980191	3000.000000	NaN	NaN	NaN	NaN	NaN	NaN
0	NaN	NaN	NaN	NaN	0.993138	0.985573	0.965907	0.965514	0.931814	0.945356

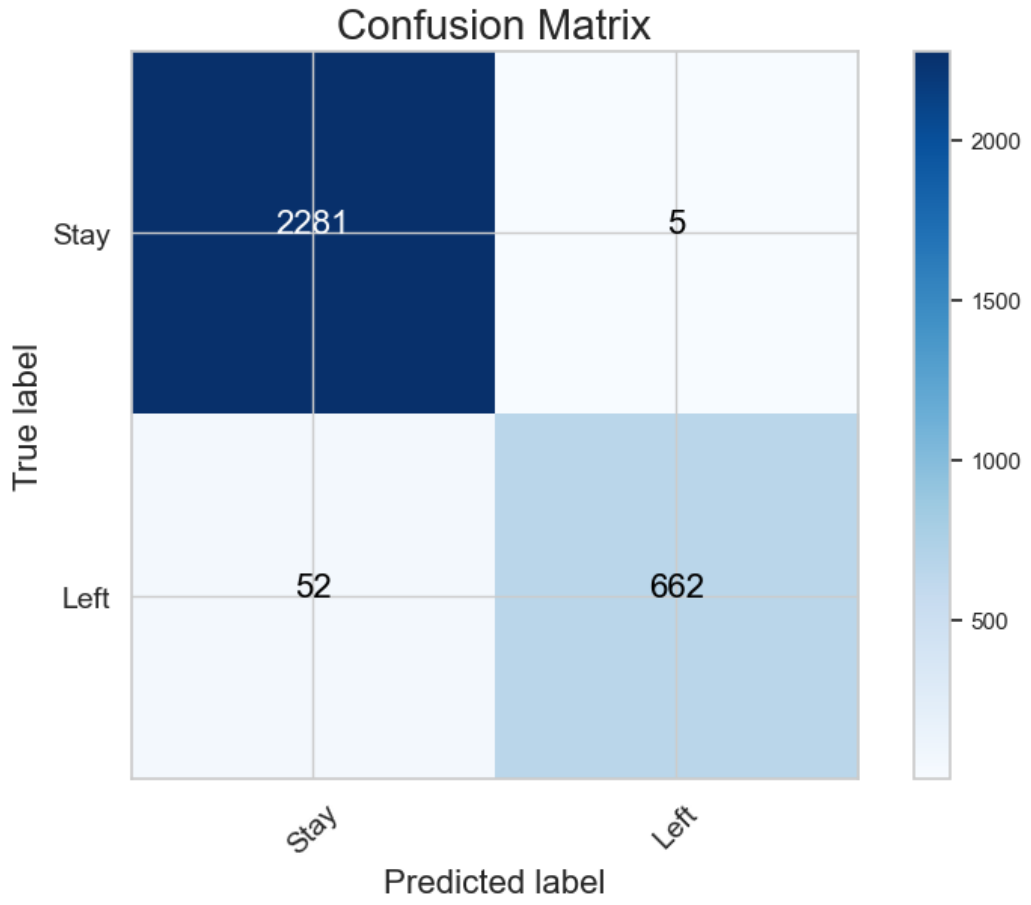
LightGBM



LGMB Classification Report:

	precision	recall	f1-score	support	ROC AUC	AP	Balanced accuracy	G-mean	Youden's index	MCC
Stay	0.981857	0.994313	0.988046	2286.000000	NaN	NaN	NaN	NaN	NaN	NaN
Left	0.981022	0.941176	0.960686	714.000000	NaN	NaN	NaN	NaN	NaN	NaN
accuracy	0.981667	0.981667	0.981667	0.981667	NaN	NaN	NaN	NaN	NaN	NaN
macro avg	0.981440	0.967745	0.974366	3000.000000	NaN	NaN	NaN	NaN	NaN	NaN
weighted avg	0.981659	0.981667	0.981534	3000.000000	NaN	NaN	NaN	NaN	NaN	NaN
0	NaN	NaN	NaN	NaN	0.993138	0.985573	0.967745	0.96738	0.93549	0.949086

Random Forest



RF Classification Report:

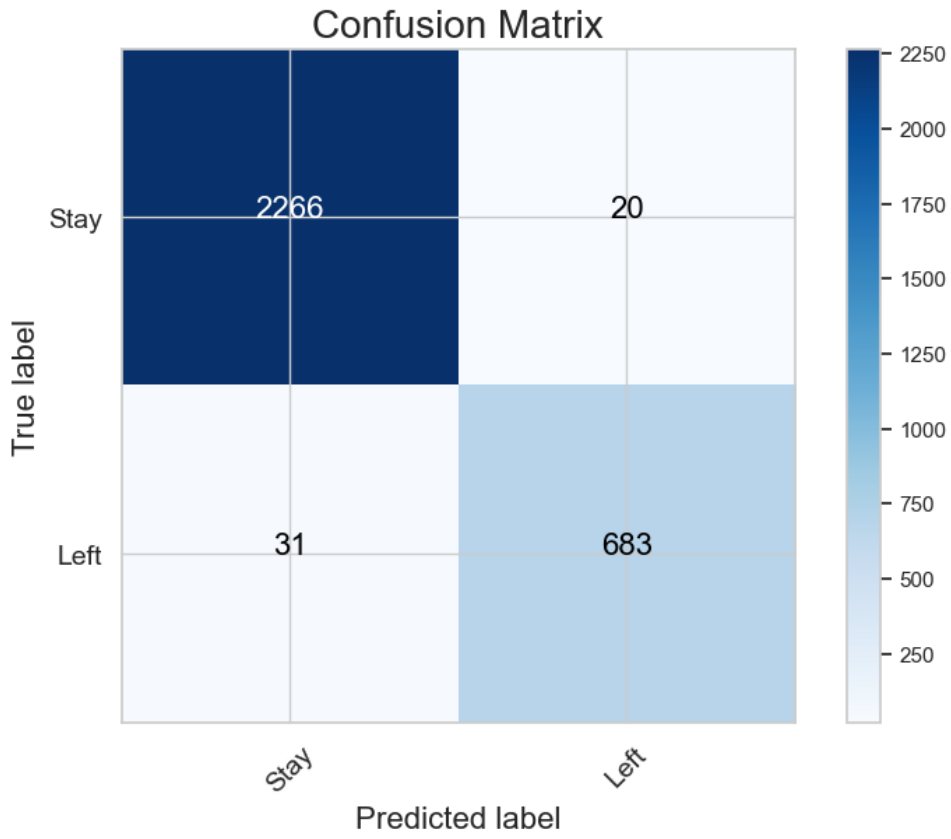
	precision	recall	f1-score	support	ROC AUC	AP	Balanced accuracy	G-mean	Youden's index	MCC
Stay	0.977711	0.997813	0.987660	2286.000	NaN	NaN	NaN	NaN	NaN	NaN
Left	0.992504	0.927171	0.958726	714.000	NaN	NaN	NaN	NaN	NaN	NaN
accuracy	0.981000	0.981000	0.981000	0.981	NaN	NaN	NaN	NaN	NaN	NaN
macro avg	0.985107	0.962492	0.973193	3000.000	NaN	NaN	NaN	NaN	NaN	NaN
weighted avg	0.981232	0.981000	0.980773	3000.000	NaN	NaN	NaN	NaN	NaN	NaN
0	NaN	NaN	NaN	NaN	0.99077	0.981834	0.962492	0.961844	0.924984	0.947329

Comparing the models

	Model	F1-score	Accuracy	ROC AUC	Balanced accuracy	Geometric mean (G-mean)	Youden's index	Matthew's correlation coefficient (MCC)	Training_time (seconds)
2	LightGBM	0.974366	0.985573	0.993138	0.967745	0.967380	0.935490	0.949086	0.099514
3	Random Forest	0.973193	0.981834	0.990770	0.962492	0.961844	0.924984	0.947329	0.233040
1	XGBoost	0.972502	0.985573	0.993138	0.965907	0.965514	0.931814	0.945356	0.087067
0	Logistic Regression	0.802108	0.724667	0.781394	0.716282	0.716103	0.432564	0.381403	0.902971

Fine-tuning LGBM

```
LGBMClassifier
LGBMClassifier(class_weight={0: 0.66, 1: 2.1}, colsample_bytree=1,
               learning_rate=0.05, max_depth=15, min_child_weight=0.01,
               n_estimators=250, n_jobs=-1, num_leaves=35, random_state=42,
               reg_alpha=0.1, reg_lambda=0.1, subsample=0.8)
```



```
Classification report:
              precision    recall  f1-score   support

        Stay     0.9865    0.9913    0.9889      2286
        Left     0.9716    0.9566    0.9640       714

    accuracy                         0.9830      3000
   macro avg     0.9790    0.9739    0.9764      3000
weighted avg     0.9829    0.9830    0.9830      3000

Area Under the Receiver Operating Characteristic Curve (ROC AUC): 0.99239
Average precision (AP)                                           : 0.98561
Balanced accuracy                      : 0.97392
Geometric mean (G-mean)                : 0.97376
Youden's index                         : 0.94783
Matthew’s correlation coefficient (MCC): 0.95293
```

Train vs Test set

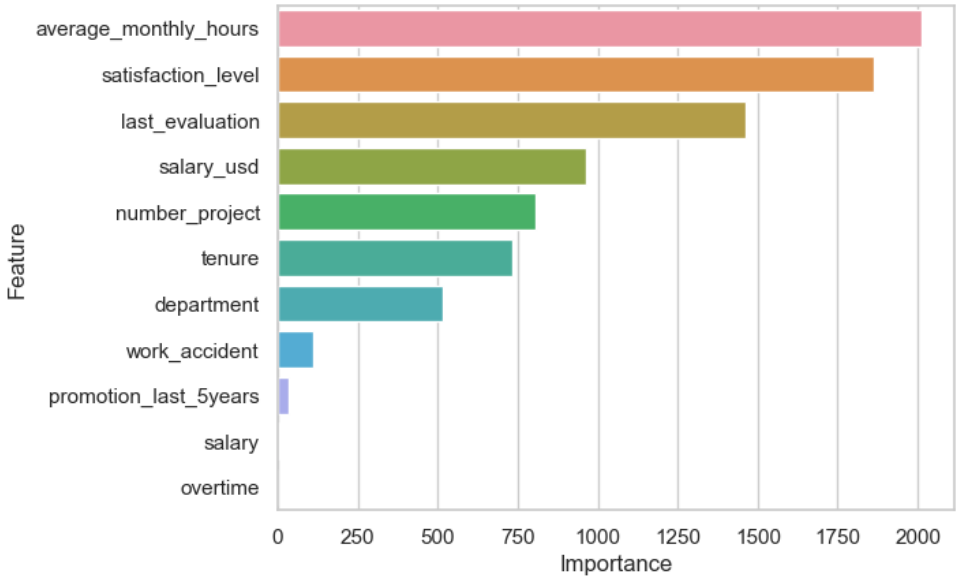
Model	Training set	Test set
Area Under the Receiver Operating Characteristic Curve (ROC AUC)	0.999956	0.992389
Average precision (AP)	0.999854	0.985606
Balanced accuracy	0.998611	0.973917
Geometric mean (G-mean)	0.998611	0.973763
Youden's index	0.997222	0.947834
Matthew's correlation coefficient (MCC)	0.995194	0.952930

	Train Set				Test Set			
	precision	recall	f1-score	support	precision	recall	f1-score	support
Stay	0.999781	0.997922	0.998850	9142.00000	0.986504	0.991251	0.988872	2286.000
Left	0.993389	0.999300	0.996336	2857.00000	0.971550	0.956583	0.964008	714.000
accuracy	0.998250	0.998250	0.998250	0.99825	0.983000	0.983000	0.983000	0.983
macro avg	0.996585	0.998611	0.997593	11999.00000	0.979027	0.973917	0.976440	3000.000
weighted avg	0.998259	0.998250	0.998252	11999.00000	0.982945	0.983000	0.982954	3000.000

The differences between the training and test set performance metrics are relatively minor, which is a good indication that the model is generalizing well and not overfitting to the training data.

The metrics suggest a robust model with strong predictive power. However, there's always a slight decrease in most metrics from the training to the test set, which is normal as the test set represents new, unseen data for the model.

Feature importance



Conclusion:

Based on the analysis conducted, we have gained valuable insights into the factors influencing employee retention. The light gradient boosting model, after hyperparameter tuning, has demonstrated strong performance in predicting employee attrition. The evaluation metrics and classification report indicate high accuracy, precision, recall, and F1-score for both the training and test sets. This suggests that the model generalizes well and can effectively identify employees at risk of leaving the company.

Recommendations:

- Enhance Job Satisfaction: The company should continuously gauge and improve job satisfaction through feedback and well-being initiatives.

- Competitive Compensation: Maintain industry-competitive salaries and offer performance incentives to keep and attract skilled employees.

- Data-Driven Decisions: Expand data collection to refine the turnover prediction model and highlight actionable retention strategies.

- Workload Oversight: Monitor and adjust employee workloads to prevent burnout, ensuring that overtime is both fair and compensated.

- Work-Life Harmony: Implement flexible and innovative work arrangements to bolster work-life balance, such as hybrid or reduced-hour schedules.

- Utilize Analytics: Use predictive analytics in conjunction with HR tools for proactive retention management.

- Invest in Development: Prioritize training programs to boost proficiency and job satisfaction, mitigating turnover.