Employee Attrition prediction

Machine Learning with Python

Context

Enhancing Retention Strategies through Employee Attrition Prediction

Employee turnover presents challenges for organizations, impacting costs and productivity.

This project aims to develop a predictive model to forecast employee attrition. Leveraging historical data on demographics, job factors, and performance metrics, the model will identify patterns and predictors of attrition.

By preemptively identifying at-risk employees, organizations can implement targeted retention strategies, fostering satisfaction and reducing talent loss. This predictive approach enables proactive HR interventions, ultimately enhancing organizational stability and success.

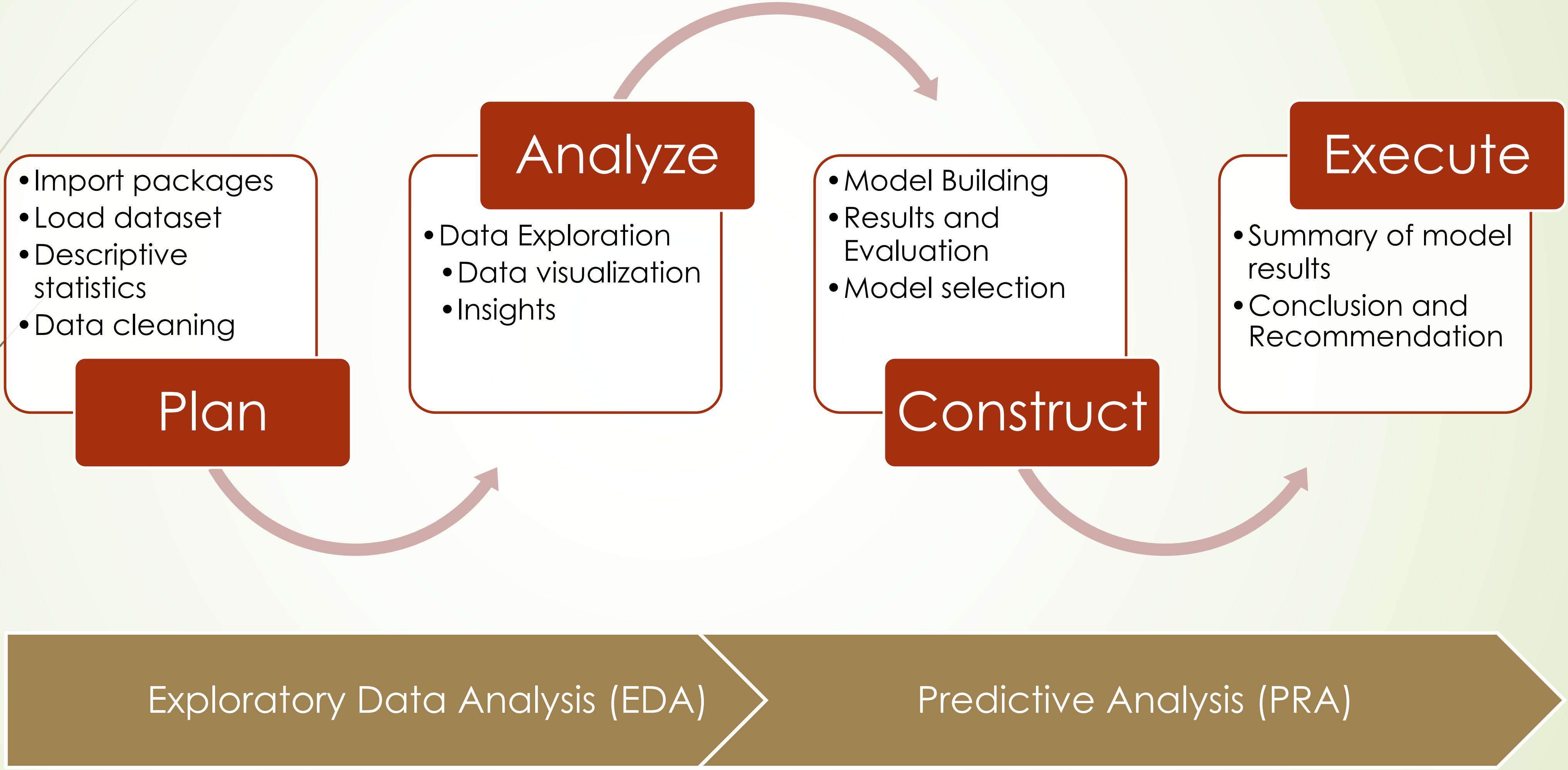
Data Description

The dataset using in this exercise contains 15,000 rows and 10 columns for the variables listed below.

For more information about the data, refer to its source on Kaggle.

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)

PACE stages approach



Plan

- Import packages
- Load dataset
- Descriptive statistics
- Data cleaning

Analyze

- Data Exploration
  - Data visualization
  - Insights

Construct

- Model Building
- Results and Evaluation
- Model selection

Execute

- Summary of model results
- Conclusion and Recommendation

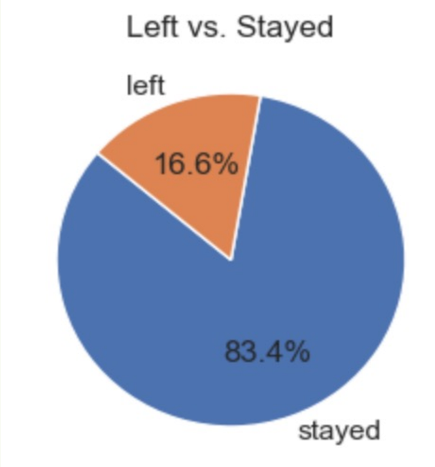
Exploratory Data Analysis (EDA)

Predictive Analysis (PRA)

Exploratory Data Analysis (EDA)

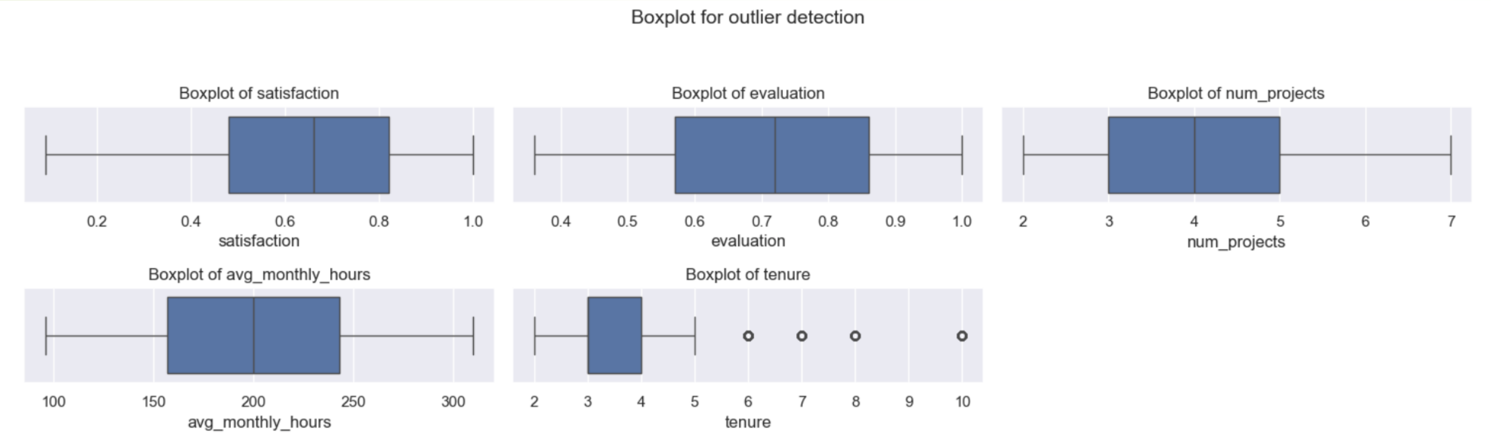
"What going on ?"

Descriptive Statistics



Analyse data set of 14,998 rows and 10 columns, we found:

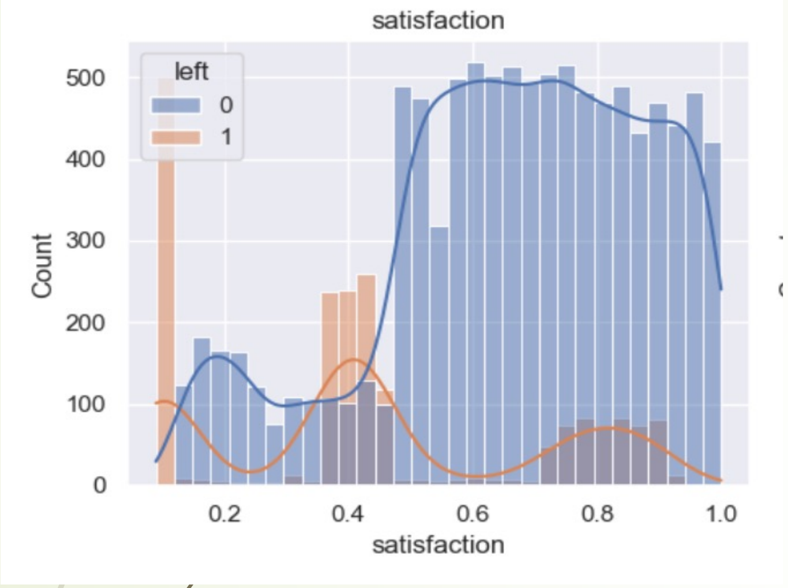
- No missing values
- 3,008 duplicated rows
- Colunm names is not standardised.
- Have outliers in ‘tenure’ variable, where the values equal and greater than 6.
- Moderately imbalanced with 16,6% ‘left’ status



Cleaning the data set:

- Removed duplicated rows.
- Standardise columns name as in snake_case format.
- Noted outliers for futher analysis when building model.
- Noted imbalanced status for statified spliting data

Statisfaction and Overworked

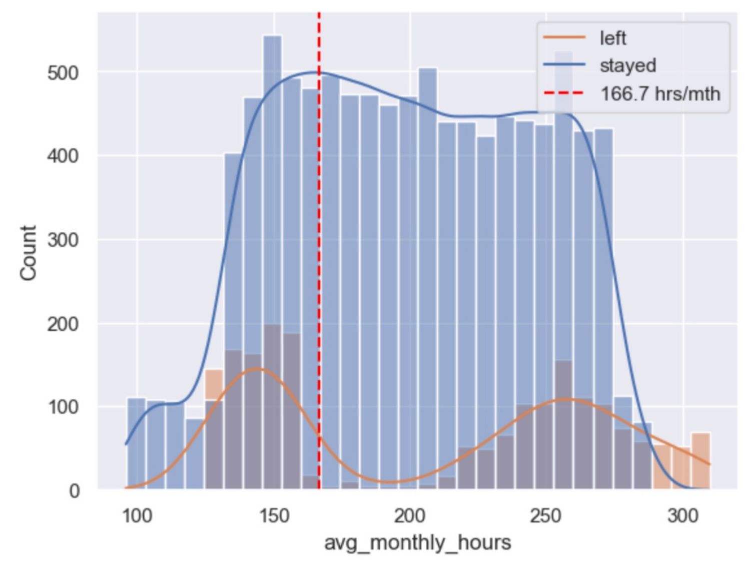


- 501 employees with satisfaction < 0.12 have left.
  - 500 in this group have not been promoted in last 5 years.
  - With total of 1991 left employees, subgroup constitutes one-fourth, suggesting that it should be separated for further analysis.
- The distribution also indicates that employees are:
  - Likely to leave when their satisfaction level is between 0.36 and 0.46.
  - Likely to stay when their satisfaction level is higher than 0.46.

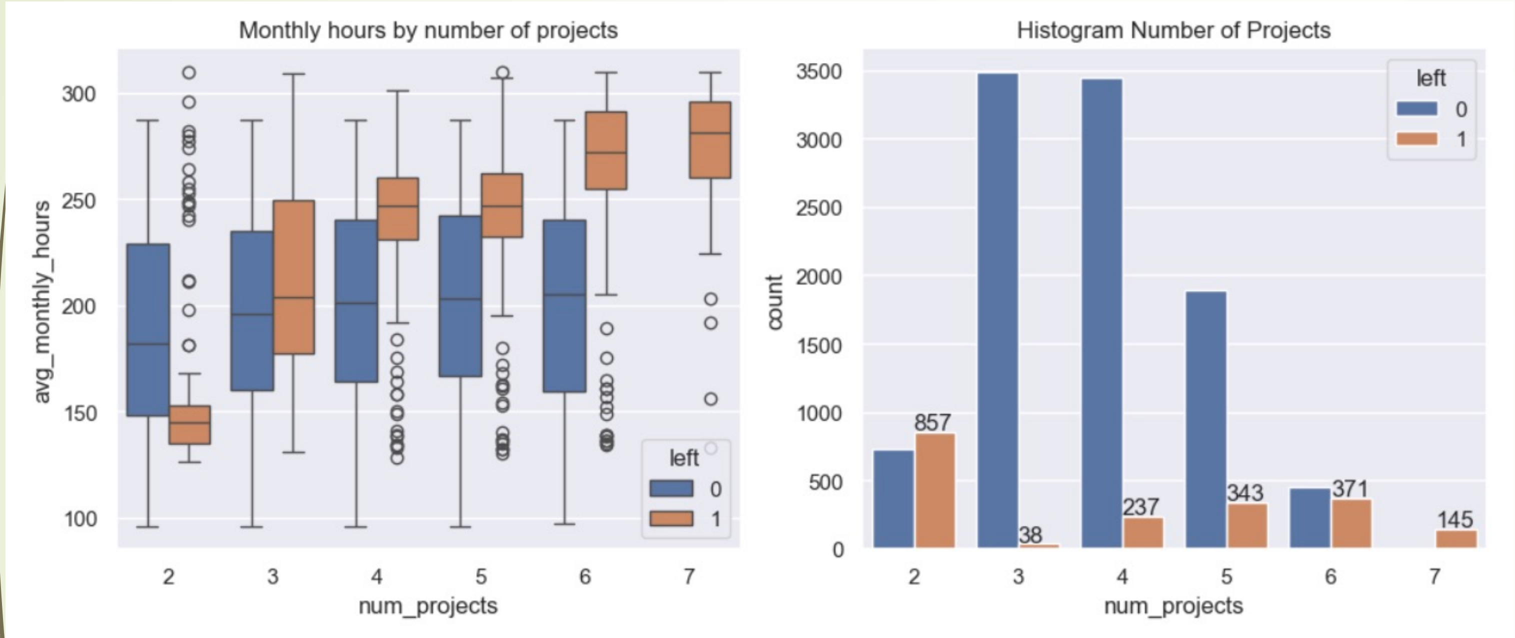
Given the Average working hours per month (‘***AVG_WK***’) is calculated as below:

$$\text{Average working hours per month} = \left(\frac{40 \text{ hours/week} \times 50 \text{ weeks}}{12 \text{ months}}\right) = \mathbf{166.7\ hrs/mth}$$

- More than 2/3 of employees have avg_monthly_hours higher than this standard, indicating that **employees are likely overworked.**
- There is a group of 238 employees working over 287 hours per month who are likely to leave the company.



Assignment and Overworked

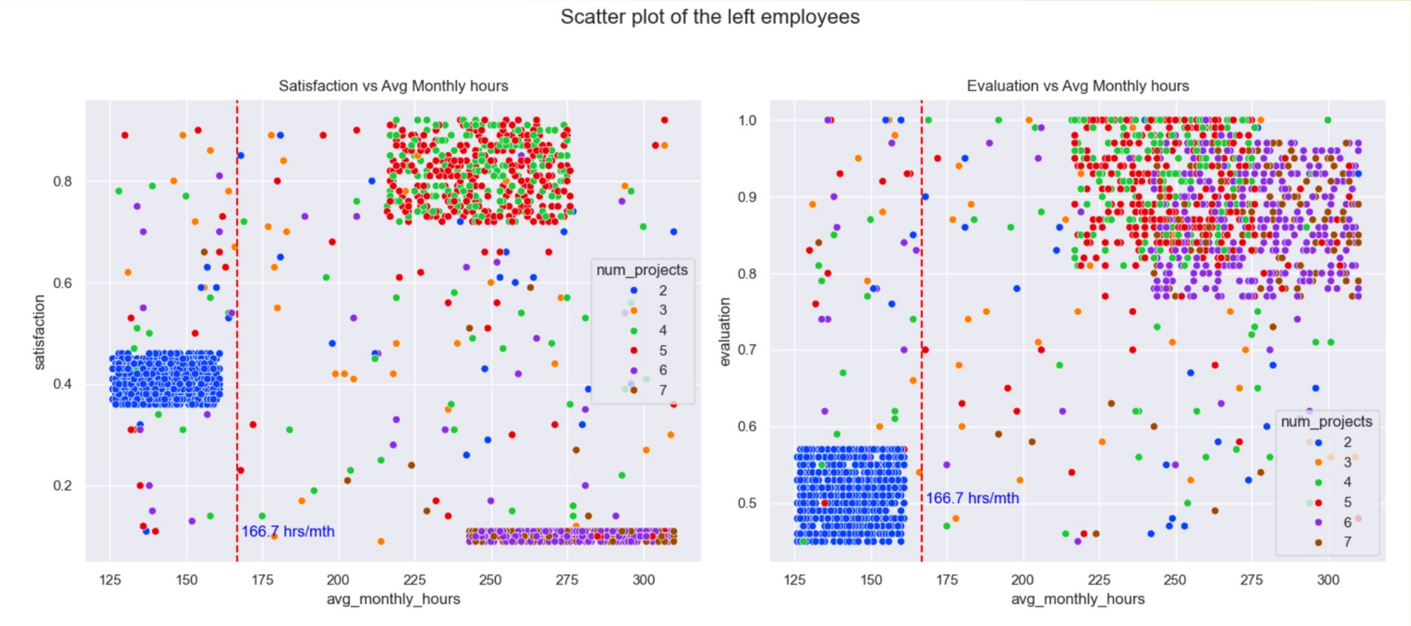


- All 145 employees working on 7 projects have left the company.
- Most of those who left while working on 6 projects having averaging 255-295 hours per month, that are higher than any other group.
- The optimal number of projects is 3 and 4, where the attrition rate is low.
- Employees working in 2 projects who left, having working hours less than others in the same number of projects

Assignment and Overworked

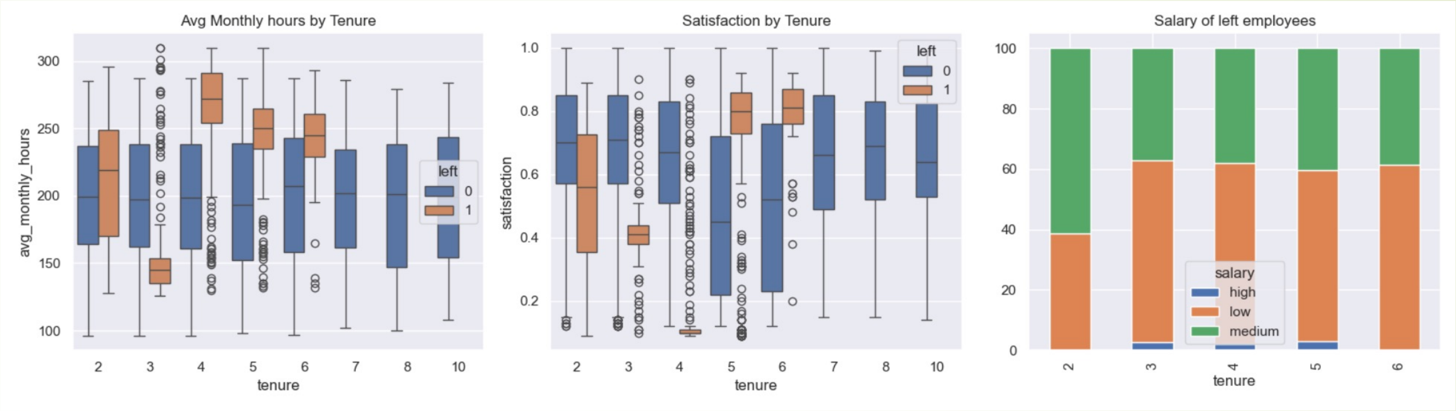
Scatter plot of employees who left the company. We can observe two groups of employees who left the company:

- (A) those work less than ***AVG_WK*** 166.7 hrs/month
- (B) those who work significantly more.



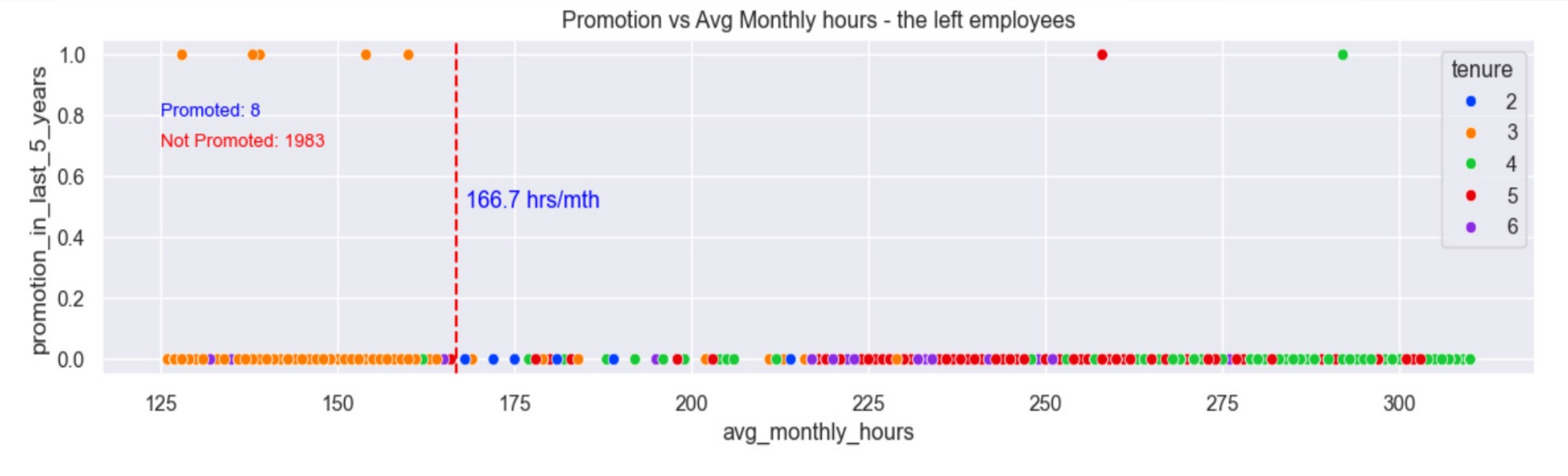
- (A): Those who worked considerably less than ***AVG_WK*** might have various reasons for their reduced workload. It's possible that some of them ***were fired or facing performance issues***, leading to a decrease in their assigned hours.
  - Their satisfaction is around 0.4, and their last evaluation score is around 0.5. (*lower than the 25th percentile of satisfaction and evaluation is 0.44 and 0.56*).
- (B) On the other hand, those who worked much more likely chose to ***quit voluntarily***. It's reasonable to infer that they probably quit due to factors such as burnout, dissatisfaction, or seeking better opportunities elsewhere. This group likely contributed significantly to the projects they worked on and might have been the largest contributors to their projects, given their high workload
  - A group of left employees working on 6 or 7 projects, with monthly hours ranging from approximately 235 to 310. They might have had good performance with high evaluation scores but ***left the company due to dissatisfaction***.
  - Another group of left employees worked on 4 or 5 projects, with workloads higher than their peers in the same number of projects. Their last evaluation scores are high, ranging from 0.8 to 1.0, but they still quit the company even though their satisfaction was above 0.7. They ***might have left for better opportunities elsewhere***.

Tenure Analysis



- Employees who have been with the company for **over 7 years are likely to stay**. There are 282 such employees as calculated above.
- Employees who have been with the company for **4, 5, or 6 years and have an average monthly hour higher than their peers in the same years are likely to leave**.
  - Employees who leave in years **5 and 6 tend to express higher dissatisfaction** compared to those who stay.
- Further examination of the salary of left employees reveals that there are no employees with high salaries in year 6, and **the majority receive low to medium salaries in years 4 and 5**.

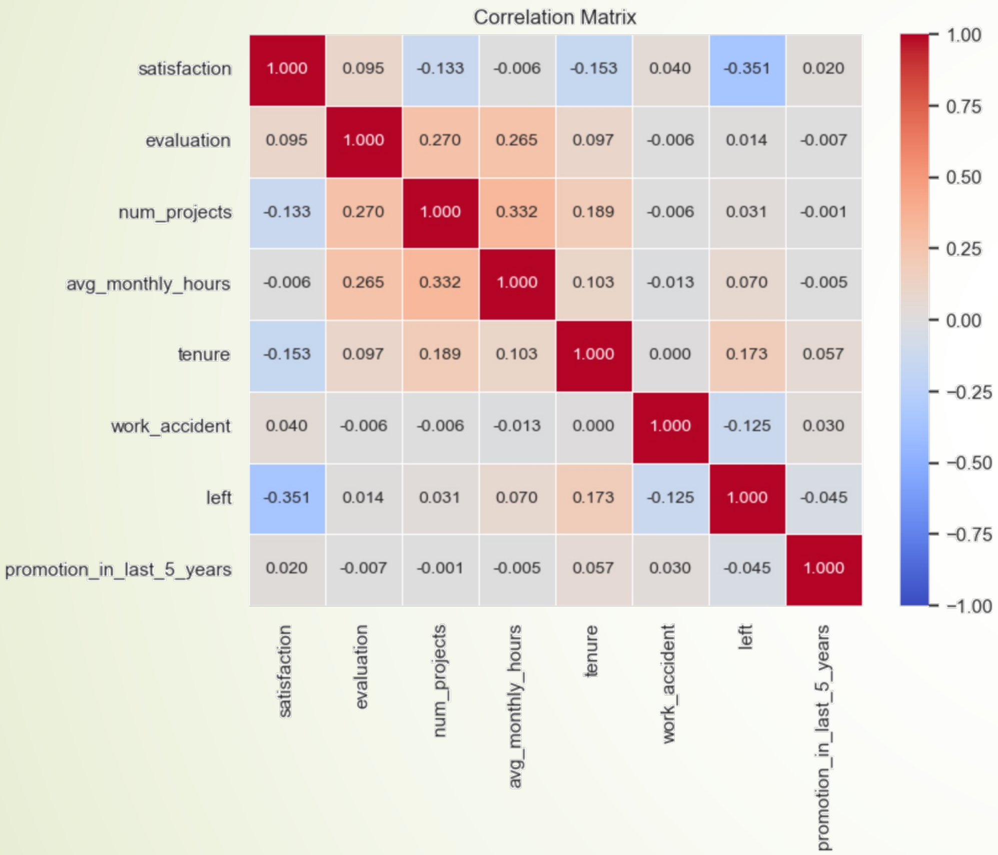
Tenue with Promotion



Within group of the left employees, we found:

- Only **8 employees have been promoted** in last 5 years !
- Employees who have been **working for 4 to 5 years and are overworked but have not been promoted yet are likely to leave**.

Correlation analysis



Correlation Matrix

	satisfaction	evaluation	num_projects	avg_monthly_hours	tenure	work_accident	left	promotion_in_last_5_years
satisfaction	1.000	0.095	-0.133	-0.006	-0.153	0.040	-0.351	0.020
evaluation	0.095	1.000	0.270	0.265	0.097	-0.006	0.014	-0.007
num_projects	-0.133	0.270	1.000	0.332	0.189	-0.006	0.031	-0.001
avg_monthly_hours	-0.006	0.265	0.332	1.000	0.103	-0.013	0.070	-0.005
tenure	-0.153	0.097	0.189	0.103	1.000	0.000	0.173	0.057
work_accident	0.040	-0.006	-0.006	-0.013	0.000	1.000	-0.125	0.030
left	-0.351	0.014	0.031	0.070	0.173	-0.125	1.000	-0.045
promotion_in_last_5_years	0.020	-0.007	-0.001	-0.005	0.057	0.030	-0.045	1.000

- **Satisfaction and Attrition**: There is a moderately strong negative correlation (-0.35) between satisfaction and the likelihood of an employee leaving the company. This implies that as satisfaction decreases, the probability of an employee leaving increases.
- **Evaluation:** There is a very weak positive correlation (0.01) between evaluation scores and the likelihood of an employee leaving. This suggests that there is not a strong linear relationship between performance evaluation scores and attrition.
- **Number of Projects**: There is a weak positive correlation (0.03) between the number of projects an employee is involved in and the likelihood of them leaving. This implies that there may be a slight tendency for employees with more projects to leave, but the effect is not very strong.
- **Average Monthly Hours**: There is a weak positive correlation (0.07) between the average monthly hours worked and the likelihood of an employee leaving. This suggests that employees who work longer hours might be slightly more likely to leave, but again, the effect is not very strong.
- **Tenure**: There is a moderate positive correlation (0.17) between the number of years an employee has been with the company and the likelihood of them leaving. This indicates that longer-tenured employees are somewhat more likely to leave compared to newer employees.
- **Work Accident and Promotion**: There are very weak correlations between work accidents, promotions in the last 5 years, and the likelihood of an employee leaving (-0.13 and -0.04, respectively). These correlations suggest that these factors have minimal linear relationships with attrition.

Predictive Analysis (PRA)

“What will happen ?”

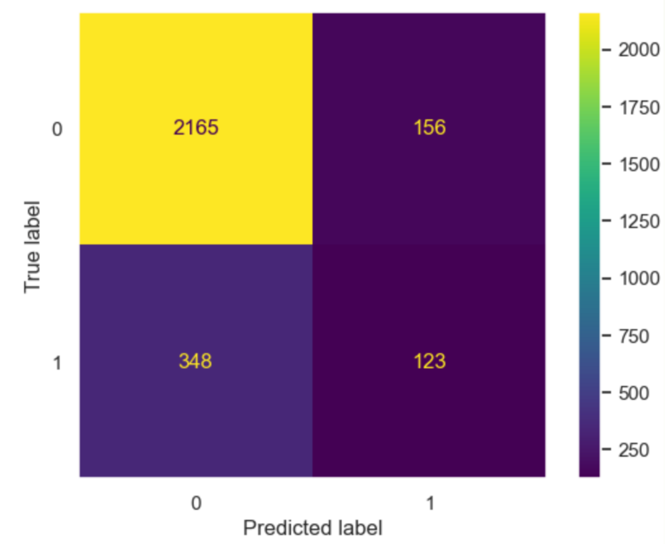
Predictive Analysis

- Our goal in this task is to predict whether an employee will leave the company or not, which is a categorical variable. So, t**his is a binary classification prediction task**.
- Since the variable we will predict (whether an employee leaves or stays) is categorical and binary, we can **use logistic regression or tree-based machine learning models for this classification task**.
- To assess the performance of ML models, we will evaluate their **confusion matrix** and measure the **evaluation scores**.

Model building & selection

Logistic Regression

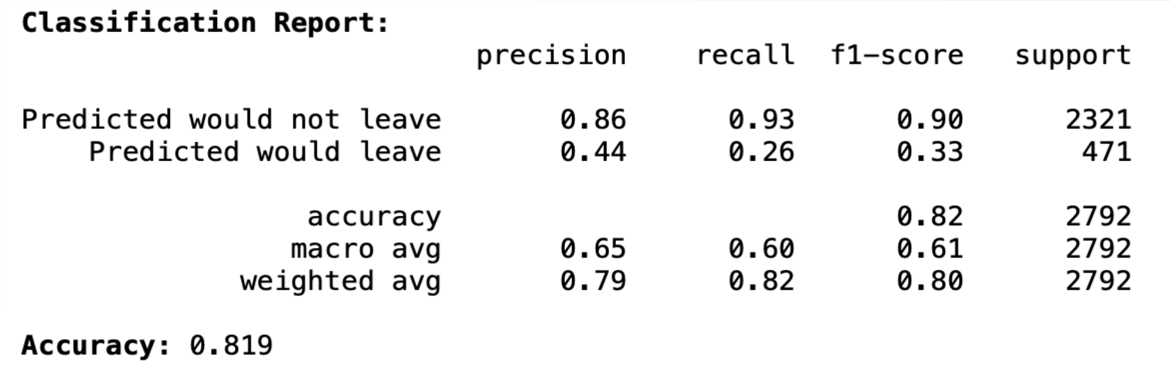
Confusion Matrix



- Since Logistic Regression (LR) is quite sensitive to outliers, we removed outliers in the tenue column before building the model. We then:
  - Constructed the LR model with the training data set,
  - Tested the model by making predictions on the test set.
- The Confusion Matrix visualizes prediction results while Evaluation metrics measure the model's performance
  - **True Negatives (TN):** 2165 employees correctly predicted to stay with the company.
  - **False Positives (FP)**: 156 employees incorrectly predicted to leave the company.
  - **False Negatives (FN)**: 348 employees incorrectly predicted to stay with the company.
  - **True Positives (TP):** 123 employees correctly predicted to leave.

Logistic Regression

Evaluation metrics



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

Predicted would not leave      0.86      0.93      0.90      2321
    Predicted would leave      0.44      0.26      0.33       471

                 accuracy                          0.82      2792
                macro avg      0.65      0.60      0.61      2792
             weighted avg      0.79      0.82      0.80      2792

Accuracy: 0.819
```

The classification report for predicting employee turnover offers key evaluation metrics:

Precision:

- "Predicted to stay": 0.86 (86% accuracy in predicting those who stay).
- "Predicted to leave": 0.44 (44% accuracy in predicting those who leave, indicating improvement needed).

Recall:

- "Predicted to stay": 0.93 (effectively captures most who stay).
- "Predicted to leave": 0.26 (misses many who actually leave).

F1-Score:

- "Predicted to stay": 0.90 (good balance between precision and recall).
- “Predicted to leave": 0.33 (low due to poor recall).

Accuracy:

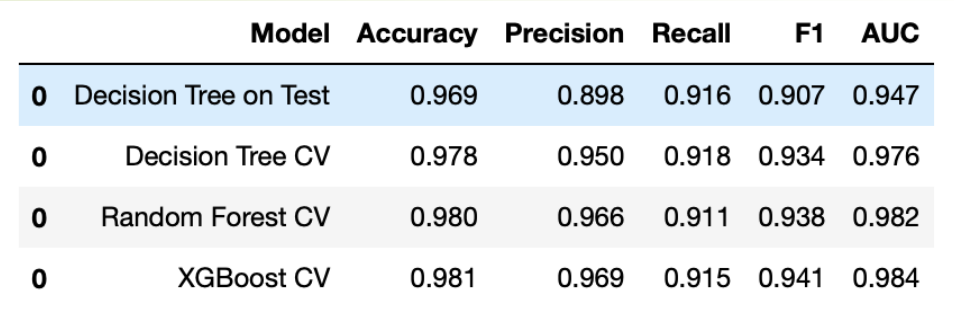
- Overall: 0.82 (82% of predictions are correct, but accuracy may not fully reflect performance on imbalanced data).

The model excels at predicting who will stay but struggles with predicting who will leave. Improving recall for the "leave" prediction could enhance overall performance, potentially through feature engineering, model tuning, or addressing class imbalance.

Tree-based models

Tuning hyperparameter & evaluating models

We continue to implement Decision Tree, Random Forest. XGBoost. After tuning hyperparameter using Cross-Validation (CV), we then evaluate the models' perfomance:



	Model	Accuracy	Precision	Recall	F1	AUC
0	Decision Tree on Test	0.969	0.898	0.916	0.907	0.947
0	Decision Tree CV	0.978	0.950	0.918	0.934	0.976
0	Random Forest CV	0.980	0.966	0.911	0.938	0.982
0	XGBoost CV	0.981	0.969	0.915	0.941	0.984

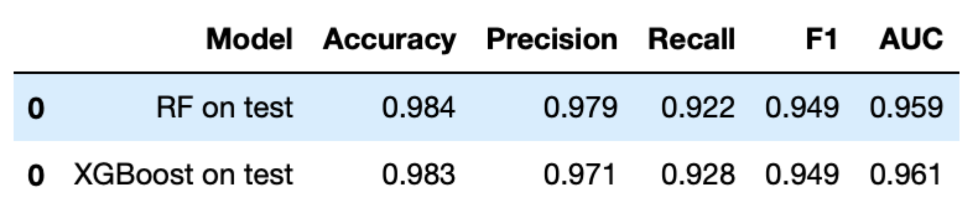
- **Decision Tree** demonstrates good performance across all metrics, with particularly high recall and AUC indicating its ability to correctly identify leaving case.
- **Decision Tree with CV** appears to improve the performance slightly compared to the decision tree model tested on a single holdout set, with higher precision and F1 score.
- **Random Forest with CV** generally performs well with high accuracy and AUC. However, it seems to have a slightly lower recall compared to the Decision Tree CV.
- **XGBoost with CV**, a gradient boosting algorithm, demonstrates the highest overall performance with the highest accuracy, precision, and F1 score among all models. However, its recall is slightly lower compared to the Decision Tree CV.

XGBoost CV shows the best overall performance across all metrics, followed closely by Random Forest CV and Decision Tree CV.

Tree-based models

Evaluating models’ performance on test set

Next, we evaluate XGBoost with CV and Random Forest with CV on test set:



	Model	Accuracy	Precision	Recall	F1	AUC
0	RF on test	0.984	0.979	0.922	0.949	0.959
0	XGBoost on test	0.983	0.971	0.928	0.949	0.961

- **Accuracy:** Random Forest with CV slightly outperforms XGBoost with CV by a margin of 0.001.
- **Precision**: Random Forest with CV achieves higher precision 0.979 compared to XGBoost with CV 0.971, indicating better identification of true positives with fewer false positives.
- **Recall:** XGBoost with CV has a higher recall of 0.928 compared to Random Forest with CV 0.922, indicating its ability to capture more true positives while possibly accepting more false positives.
- **F1 Score:** Both models have the same F1 score of 0.949, which is the harmonic mean of precision and recall, suggesting that both models achieve a good balance between precision and recall.
- **AUC:** XGBoost has a slightly higher AUC of 0.961, compared to Random Forest's AUC of 0.959. AUC, representing the model's ability to distinguish between positive and negative instances. In this case, XGBoost has a slightly better ability to do so..

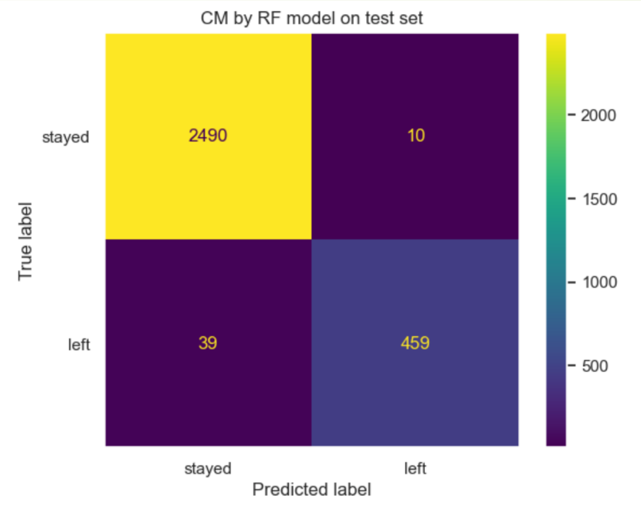
Both Random Forest and XGBoost models perform exceptionally well in predicting employee attrition on the test set.

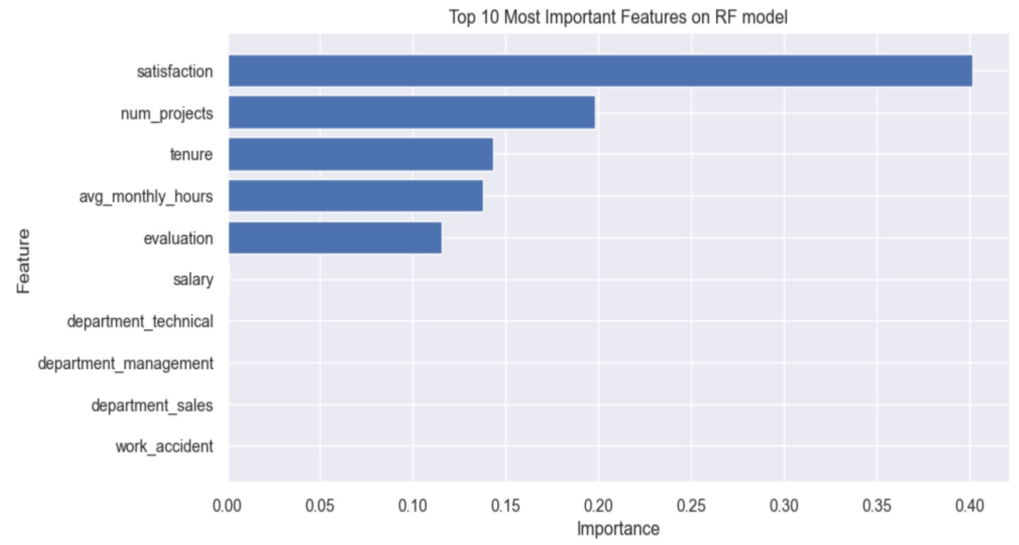
Random Forest performs slightly better in terms of accuracy, precision, and AUC, while XGBoost shows a slightly higher recall. Giving these results, Random Forest is suggested as selected model for prediction.

Random Forest

Confusion Matrix & Feature Importances

- The confusion matrix plot illustrates how well the Random Forest model predicts on the test set:
  - The lower number of false negatives (39) and false positives (10) highlights the model's effectiveness in predicting employee attrition, demonstrating its ability to accurately identify employees who are likely to leave the company.





- The feature importances plot highlights that ***satisfaction level***, ***number of projects***, ***tenure***, ***average monthly hours***, and ***evaluation score*** have the highest importance scores. These variables are most helpful in predicting the outcome variable '***left***', aligning with the key factors identified during the exploratory analysis steps.

Summary of Models results

	Accuracy	Precision	Recall	F1
Logistic Regression	82%	79%	82%	80%
Decision Tree	96.9%	89.6%	91.6%	90.7%
Random Forest	98.4%	97.9%	92.2%	94.9%
XGBoost	98.3%	97.1%	92.8%	94.9%

Based on the model performance metrics, **Random Forest with Cross Validation** appears to be the most appropriate model due to its consistently high accuracy (0.984), precision (0.979), recall (0.922), F1 score (0.949), and AUC (0.959), indicating robust predictive capability across various evaluation metrics.

Conclusion and Recommendations

Based on the exploratory and predictive analysis, it is confirmed that employees are overworked and leave the company due to dissatisfaction.

To retain employees, the following recommendations could be implemented:

Enhance Employee Satisfaction

- Focus on improving overall satisfaction, especially for employees with satisfaction levels between 0.36 and 0.46, to reduce turnover.
- Implement measures to identify and address dissatisfaction early, particularly for employees below the 0.12 satisfaction threshold.

Promotion and Career Development

- Establish clear career progression and promotion pathways, particularly targeting employees within their 4th and 5th years.
- Recognize and promote deserving employees to boost morale and retention.

Manage Workload

- Monitor and balance workloads to prevent burnout, especially for employees working over 287 hours per month.
- Aim to keep project involvement at an optimal level (3-4 projects) to maintain productivity without causing excessive stress.

Address Tenure-Related Issues

- Provide additional support and incentives for employees with 4-6 years of tenure to improve retention.
- Implement retention strategies for mid-tenure employees, including competitive salaries and professional development opportunities.

Focus on Compensation

- Ensure competitive compensation, particularly for employees with medium to low salaries in their 4th and 5th years.
- Regularly review and adjust salary structures to retain high-performing employees.

Prediction and Prevention

- The selected Random Forest model with Cross Validation demonstrates strong performance across all metrics, indicating its effectiveness in predicting and preventing employee leaving the company. By leveraging this model, organizations can proactively identify employees at risk of leaving and implement targeted retention strategies to improve employee retention rates and maintain workforce stability.

Thank you