SWE 599

Employee Churn Prediction

Oğuz Senna
Advisor: Berk Gökberk

October 22, 2023

Table of Contents

1. Plan:

Task	**Deadline**
Plan	**October 22, 2023**
Progress Report	**November 26, 2023**
Final Project	**June 3, 2024**

1. Plan:

For the SWE 599 course, I have chosen to implement the "Employee Churn Prediction" project, building upon concepts learned in SWE 578. The primary objective is to develop an effective model for predicting employee attrition within an organization. To achieve this, I will employ a range of popular machine learning and artificial intelligence methodologies including logistic regression, XGBoost, and Random Forest.

The utilization of these diverse algorithms will provide a comprehensive approach to address the complexity of the churn prediction problem. For data acquisition, I will leverage reputable sources on Kaggle, ensuring access to high-quality datasets with real-world relevance. The project plan encompasses data preprocessing, feature engineering, model selection, training, validation, and thorough evaluation.

Additionally, I will implement techniques for model interpretability and explore potential areas for improvement. This project will not only showcase a solid understanding of SWE 578 concepts but also demonstrate practical application and critical thinking in the domain of employee churn prediction.

Project Codes:

https://colab.research.google.com/drive/1e4jq6RTHsRyN41nHQRgPyLehKw86hxeP?usp=sharing

2. Progress Report:

I did some research to find a suitable database for the model implementation. I found a database that was created by IBM Data Scientists.

Dataset can be found at the link below:

https://www.kaggle.com/code/faressayah/ibm-hr-analytics-employee-attrition-performance/input

When I investigate the data, there are a total of 35 features. 9 of them are “object” type and 26 of them are “int64” type. Also, there are 1470 columns of employee data in the dataset.

2.a. Dataset:

- **Age**	1470 non-null	int64
- **Attrition**	1470 non-null	object
- **BusinessTravel**	1470 non-null	object
- **DailyRate**	1470 non-null	int64
- **Department**	1470 non-null	object
- **DistanceFromHome**	1470 non-null	int64
- **Education**	1470 non-null	int64
- **EducationField**	1470 non-null	object
- **EmployeeCount**	1470 non-null	int64
- **EmployeeNumber**	1470 non-null	int64
- **EnvironmentSatisfaction**	1470 non-null	int64
- **Gender**	1470 non-null	object
- **HourlyRate**	1470 non-null	int64
- **JobInvolvement**	1470 non-null	int64
- **JobLevel**	1470 non-null	int64
- **JobRole**	1470 non-null	object
- **JobSatisfaction**	1470 non-null	int64
- **MaritalStatus**	1470 non-null	object
- **MonthlyIncome**	1470 non-null	int64
- **MonthlyRate**	1470 non-null	int64
- **NumCompaniesWorked**	1470 non-null	int64
- **Over18**	1470 non-null	object
- **OverTime**	1470 non-null	object
- **PercentSalaryHike**	1470 non-null	int64
- **PerformanceRating**	1470 non-null	int64
- **RelationshipSatisfaction**	1470 non-null	int64
- **StandardHours**	1470 non-null	int64
- **StockOptionLevel**	1470 non-null	int64
- **TotalWorkingYears**	1470 non-null	int64
- **TrainingTimesLastYear**	1470 non-null	int64
- **WorkLifeBalance**	1470 non-null	int64
- **YearsAtCompany**	1470 non-null	int64

- **YearsInCurrentRole** 1470 non-null int64
- **YearsSinceLastPromotion** 1470 non-null int64
- **YearsWithCurrManager** 1470 non-null int64

2.b.i. Logistic Regression:

Description:

For categorization and predictive analytics, statistical models like logistic regression, sometimes referred to as the logit model, are frequently used. It uses independent factors from a given dataset to assess the likelihood of an occurrence, like voting or not. Because it expresses a probability, the dependent variable in logistic regression is bounded between 0 and 1.

Type:

Logistic regression is a form of regression analysis used to estimate the likelihood of an event occurring.

Nature of the Model:

In spite of its name, binary classification tasks—which involve estimating the likelihood that an instance will fall into a particular category—are the main application for logistic regression.

Algorithm:

By using the logistic function, logistic regression converts a linear combination of input data into a number between 0 and 1, which represents the likelihood that the instance would fall into the positive class.

Interpretability:

Logistic regression is easily interpretable, and its model coefficients provide insights into the importance of features.

Use Case in Employee Churn Prediction:

Logistic regression is suitable for modeling the probability of an employee leaving based on factors like salary, job satisfaction, and tenure.

2.b.ii. Random Forest:

Description:

A popular machine learning approach called random forest creates a single outcome by aggregating the outputs of several decision trees. It can handle problems with both regression and classification, making it flexible.

Type:

One ensemble learning technique that can be used for both regression and classification tasks is called Random Forest.

Nature of the Model:

Each decision tree in the ensemble has been trained using a different subset of the data and features. The average or vote over each individual tree forecast yields the final projection.

Algorithm:

During training, numerous decision trees are built using Random Forest, and the final prediction is decided by average or voting over these trees' predictions.

Advantages:

Robustness, good generalization, and resistance to overfitting are characteristics of Random Forests. They are good at capturing intricate data linkages.

Use Case in Employee Churn Prediction:

Because Random Forests can handle a large number of characteristics and capture non-linear relationships in the data, they are useful for predicting staff attrition.

2.b.iii. XGBoost:

Description:

XGBoost is an ensemble learning technique that extends the gradient boosting framework for better speed and performance. It is based on decision trees. It is well known for being effective and efficient, and machine learning contests frequently employ it.

Gradient Boosting:

XGBoost sequentially improves weak learners, which are mostly decision trees, with each new tree fixing mistakes committed by the ones before it.

Regularization:

To manage model complexity and avoid overfitting, XGBoost includes L1 (LASSO) and L2 (ridge) regularization factors in the objective function.

Managing Absent Values:

A built-in method for handling missing values in XGBoost improves resilience in real-world datasets where missing data is frequent.

Parallel Processing:

Because of its fast design, XGBoost may be parallelized on a distributed computer cluster and benefit from parallel processing.

Tree Pruning:

In order to avoid overfitting, XGBoost prunes trees while they are being built, eliminating splits that have little bearing on the model's functionality.

2.b.iv. Support Vector Machine (SVM):

Description:

Encouragement A class of supervised learning techniques called vector machines is applied to regression and classification problems. The main goal is to locate the hyperplane in an N-dimensional space that best divides the data into classes.

Kernel Trick:

In order to improve the algorithm's capacity to identify an efficient hyperplane, SVMs employ a "kernel trick" to convert input data into a higher-dimensional space.

Margin Maximization:

SVM seeks to maximize the margin, which indicates the classifier's confidence, between the nearest data points of each class and the decision border (hyperplane).

Support Vectors:

Support vectors, which are most important in establishing the location and orientation of the hyperplane, are the data points that are closest to the decision boundary.

Non-linearity:

By using different kernels (such as polynomial and radial basis functions) to transfer data into a higher-dimensional space, SVM can handle non-linear decision boundaries.

2.c. Data Cleaning and Preprocessing:

Age	0
Attrition	0
BusinessTravel	0
DailyRate	0
Department	0
DistanceFromHome	0
Education	0
EducationField	0
EmployeeCount	0
EmployeeNumber	0
EnvironmentSatisfaction	0
Gender	0
HourlyRate	0
JobInvolvement	0
JobLevel	0
JobRole	0
JobSatisfaction	0
MaritalStatus	0
MonthlyIncome	0
MonthlyRate	0
NumCompaniesWorked	0
Over18	0
OverTime	0
PercentSalaryHike	0
PerformanceRating	0
RelationshipSatisfaction	0
StandardHours	0
StockOptionLevel	0
TotalWorkingYears	0
TrainingTimesLastYear	0
WorkLifeBalance	0
YearsAtCompany	0
YearsInCurrentRole	0
YearsSinceLastPromotion	0
YearsWithCurrManager	0

- Data shape is (1470, 35).
- **Object type features includes**; Attrition, BusinessTravel, Department, EducationField, Gender, JobRole, MaritalStatus, Over18, OverTime.

- **Int64 type features includes**; Age, Attrition, DailyRate, DistanceFromHome, Education, EmployeeCount, EmployeeNumber, EnvironmentSatisfaction, HourlyRate, JobInvolvement, JobLevel, JobSatisfaction, MonthlyIncome, MonthlyRate, NumCompaniesWorked, PercentSalaryHike, PerformanceRating, RelationshipSatisfaction, StandardHours, StockOptionLevel, TotalWorkingYears, TrainingTimesLastYear, WorkLifeBalance, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager.

As we can see there are no missing items on the dataset, so we do not have to deal with the cleaning and processing of the data.

According to the dataset 1233 employees stayed on the other hand 237 people have left the company. By this result, %16.12 of the employees have left the company.

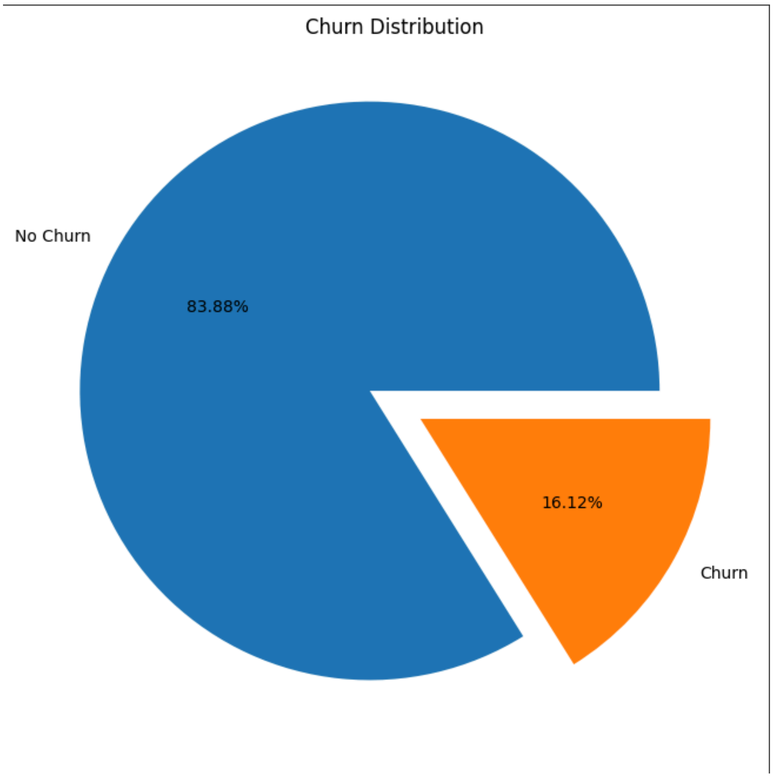

Fig 1. Churn Distribution

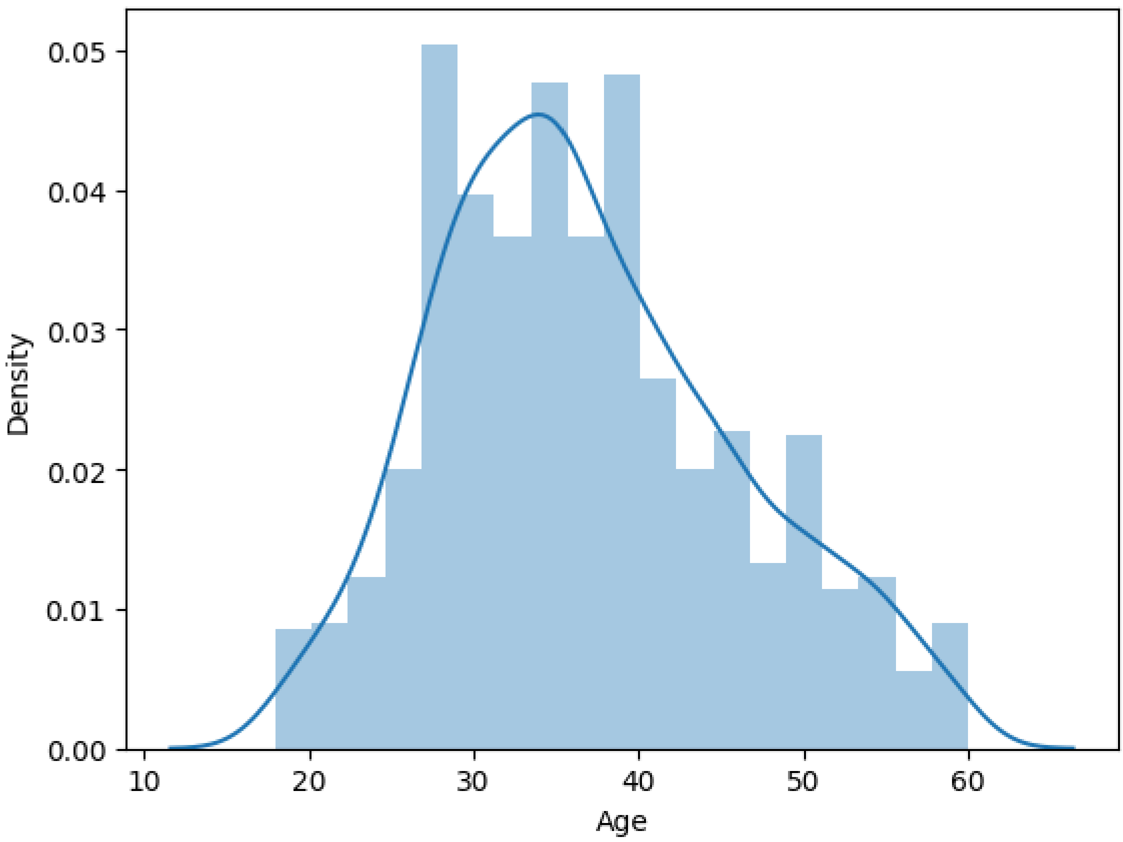

Fig 2. Age Segmentation

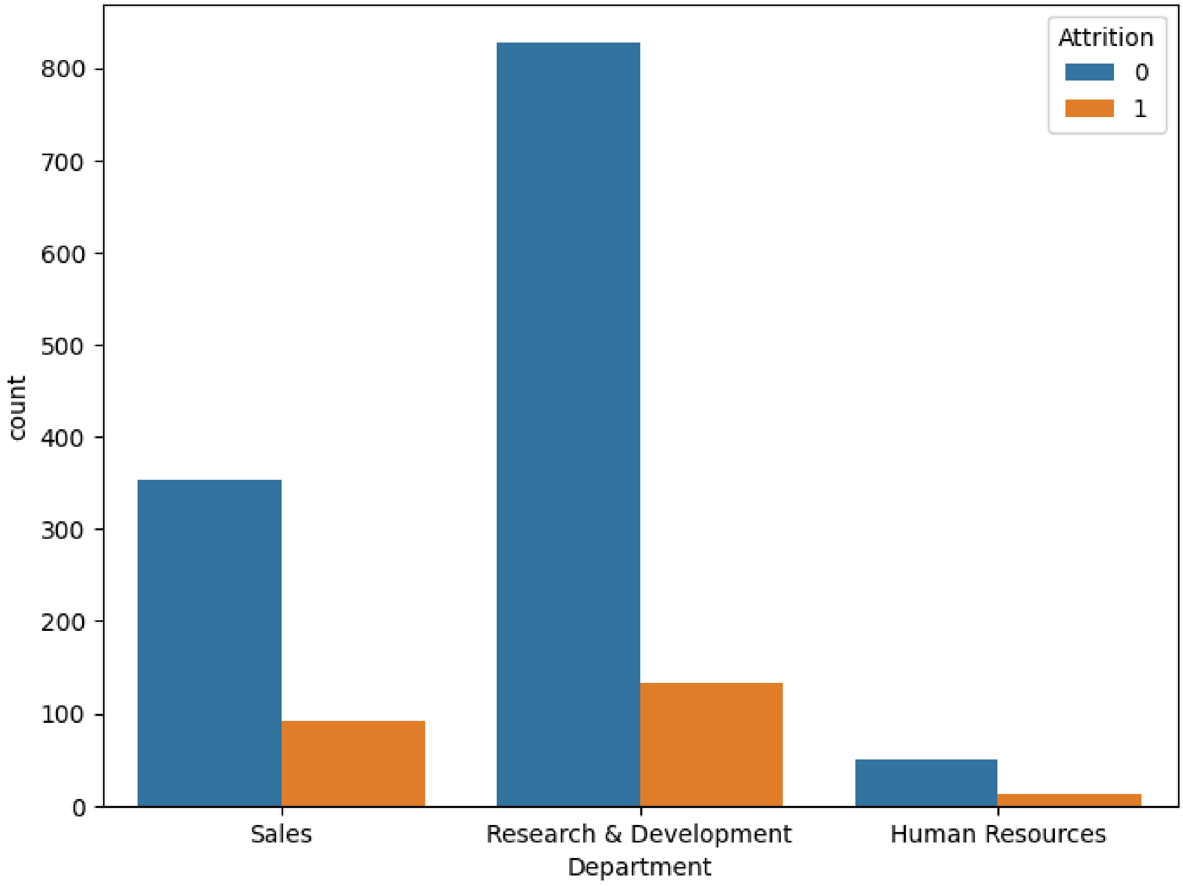

Fig 3. Department & Attrition Correlation

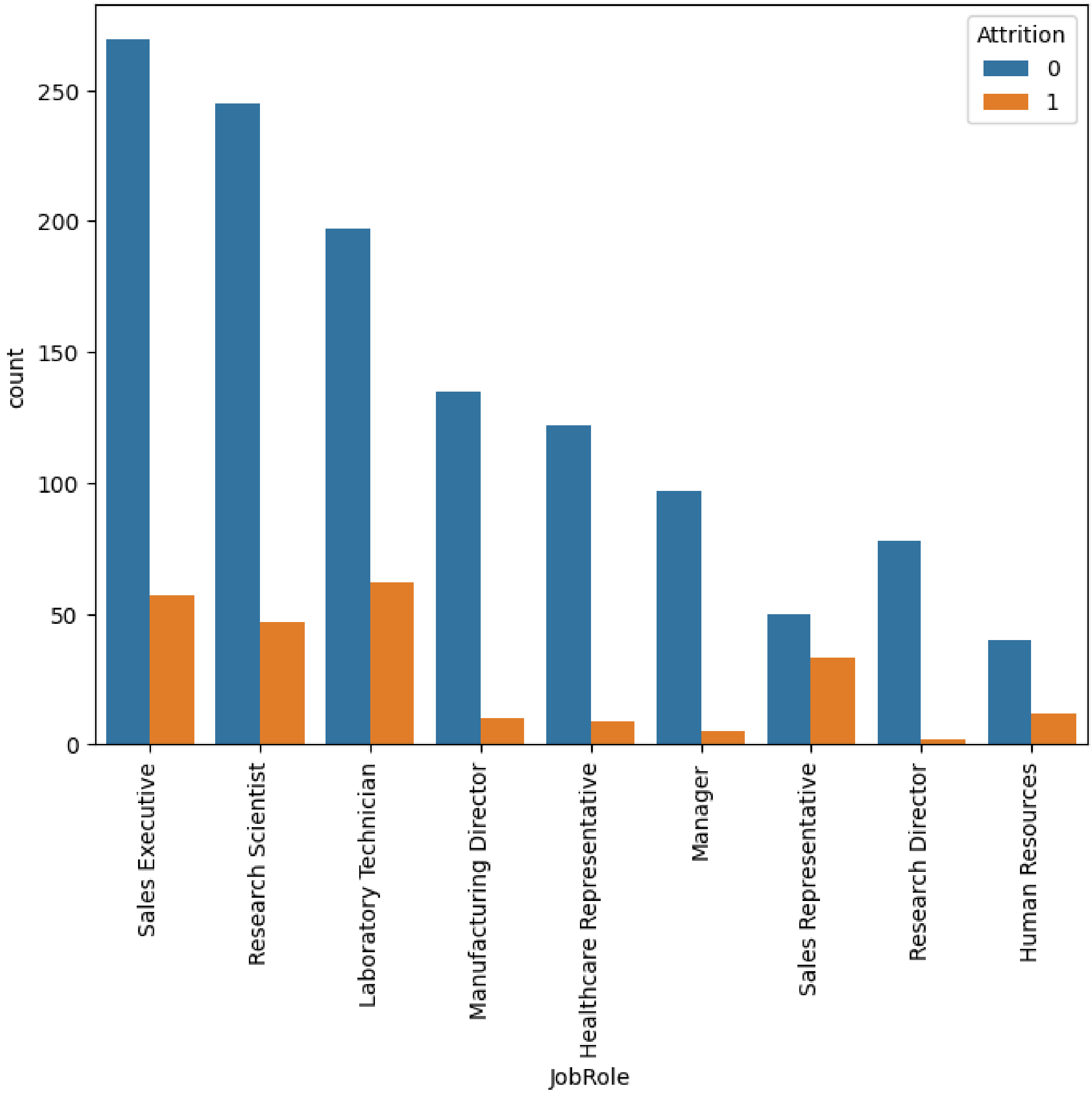

Fig 4. JobRole & Attrition Correlation

2.c.i. HeatMap:

To see the correlation analysis between the variables, we can build a heatmap.

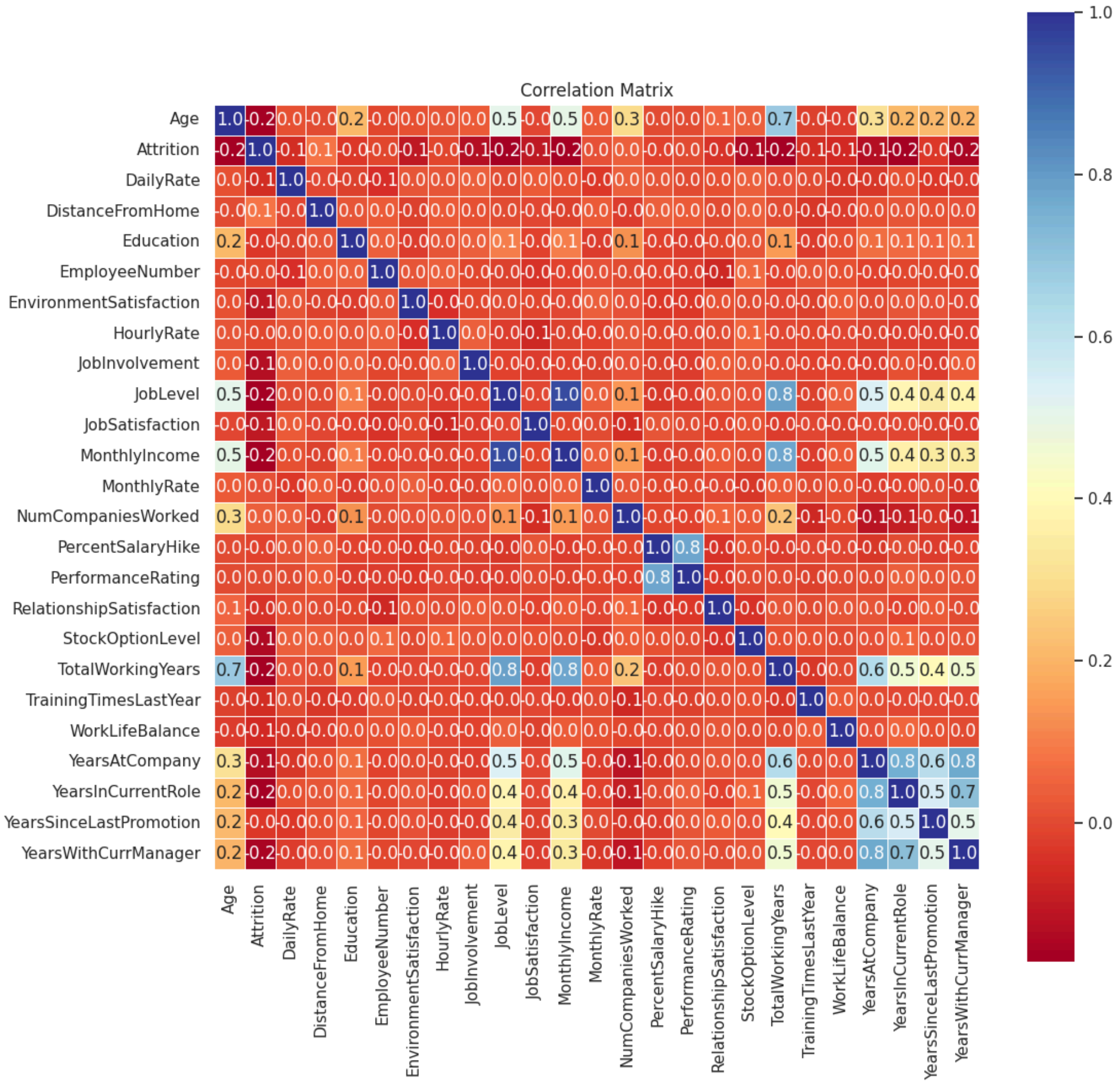

Fig 5. Heatmap for Correlation Analysis

- Senior employees tend to accumulate more total working years, which is a clear trend.
- An increase in performance ratings is associated with a higher percentage of salary hike.
- Monthly income tends to rise with the number of years an employee has been with the company.
- Many employees stay in their current roles and under the same manager over the years, indicating a lack of promotions, which could be a significant factor contributing to attrition.

2.d Data Manipulation

Firstly, we need to identify features and the target. Attrition, BusinessTravel, EducationField and OverTime seems not the key factor on the dataset. So that I will drop these features from the X variable. Target variable is the Attrition from the dataset.

We need to label encode the categorical variables before training the models. 'Department', 'Education', 'JobRole', 'Gender', 'MaritalStatus', 'Over18' these features are encoded. Afterwards, I use the StandardScaler function which is from scikit-learn library to standardize the features. By standardizing the features, you make it easier for the algorithm to converge during training and avoid certain features dominating the learning process simply because they have larger scales than others.

2.e Building Models

In this research I used Logistic Regression and Random Forest, XGBoost and Support Vector Machine.

Results:

Model	Training Accuracy	Testing Accuracy	Cross-Validation Mean	Precision	Recall	F1 Score
RF	1.0000	0.8458	0.8012	0.8166	0.8458	0.7952
LR	0.8688	0.8549	0.8292	0.8316	0.8549	0.8222
XGBoost	1.0000	0.8526	0.8275	0.8274	0.8526	0.8266
SVM	0.8814	0.8481	0.7795	0.8338	0.8481	0.7935

Logistic Regression stands out as a robust performer with a good balance between precision and recall.

Random Forest and XGBoost, while achieving high accuracies, raise concerns about potential overfitting.

SVM offers competitive performance but falls slightly behind Logistic Regression in terms of F1 score.