Predicting Employee Turnover: An Analysis of the IBM HR Dataset

Piyush Dwivedi

1. Abstract:

This study delves into the intricacies of employee attrition and performance using the IBM HR Analytics dataset. With an increasing focus on talent retention and productivity optimization, understanding the factors contributing to employee turnover and performance is paramount for organizational success. Leveraging machine learning techniques, this research investigates patterns within the dataset to identify key predictors of attrition and performance. Through rigorous analysis and interpretation of results, valuable insights are gleaned, providing actionable recommendations for HR professionals and organizational leaders.

2. Motivation:

Employee attrition poses a significant challenge for organizations across industries, leading to increased recruitment costs, loss of institutional knowledge, and decreased morale among remaining employees. Conversely, employee performance is directly linked to organizational success, productivity, and competitiveness in the market. Recognizing the critical importance of addressing these issues, this study seeks to uncover underlying factors driving employee attrition and performance using advanced analytics techniques. By shedding light on these factors, organizations can proactively implement strategies to mitigate attrition risk, enhance employee engagement, and optimize workforce performance, ultimately fostering a more resilient and thriving organizational culture.

3. Dataset:

This project utilizes the publicly available IBM HR Analytics Employee Attrition & Performance dataset from Kaggle. The dataset contains approximately 1500 employee entries with various features like Age, Business Travel, Department, work satisfaction, salary, and a target variable indicating employee attrition (left the company) or retention (remained). To ensure data quality, missing values were checked there were no missing values are present and some categorical features might have been encoded numerically for model training.

4. Data cleaning:

In the process of data cleaning following things are done:

4.1. Checking Missing Values

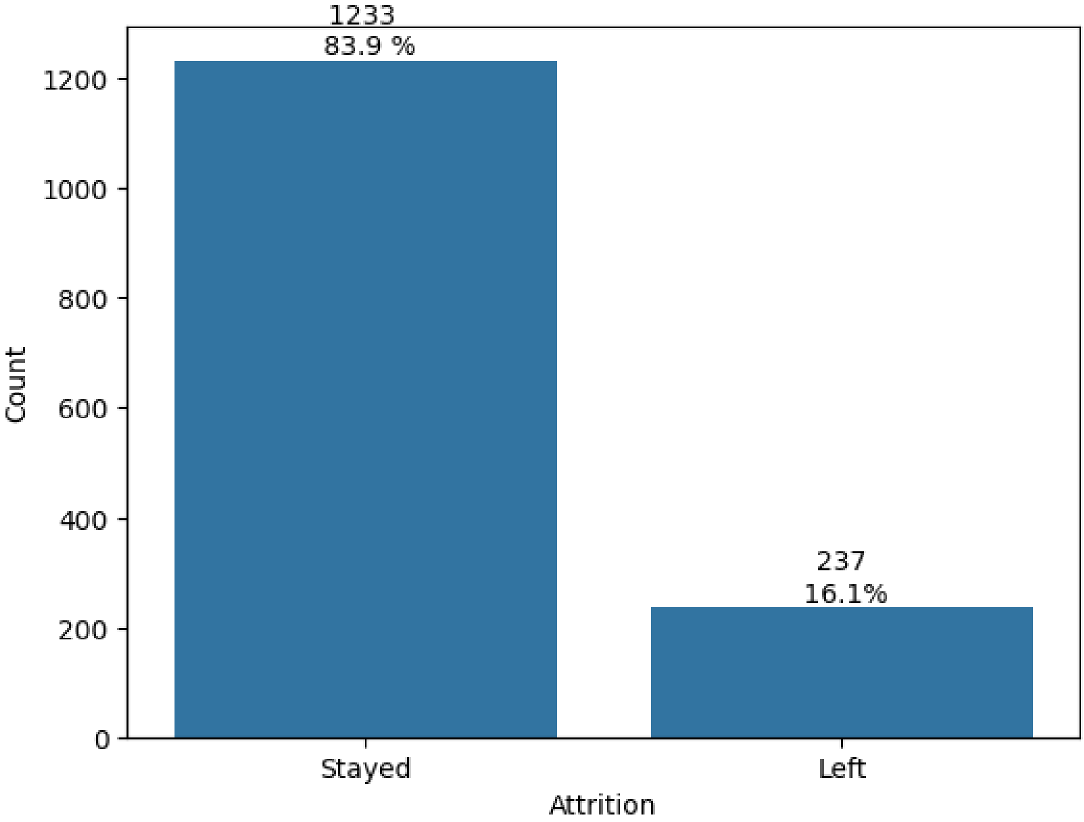
There were no missing value present in Dataset

4.2. Check duplicates

There were no duplicate value were present

4.3. Checking Imbalance in Dataset

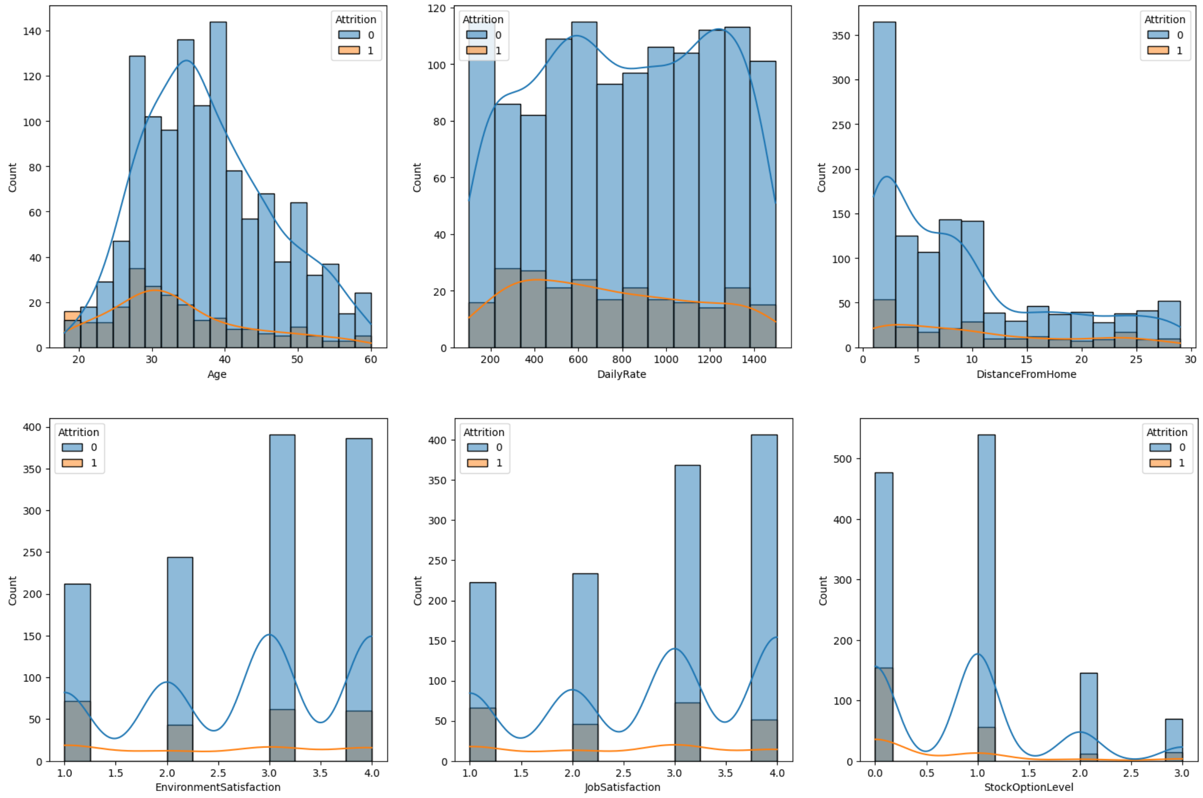
There are less people who are leaving company compare who are staying. People who are leaving company they were 16.1% and people who are no leaving they were 83.9%. This is a imbalance dataset.



5. Data Visualization & Insight's

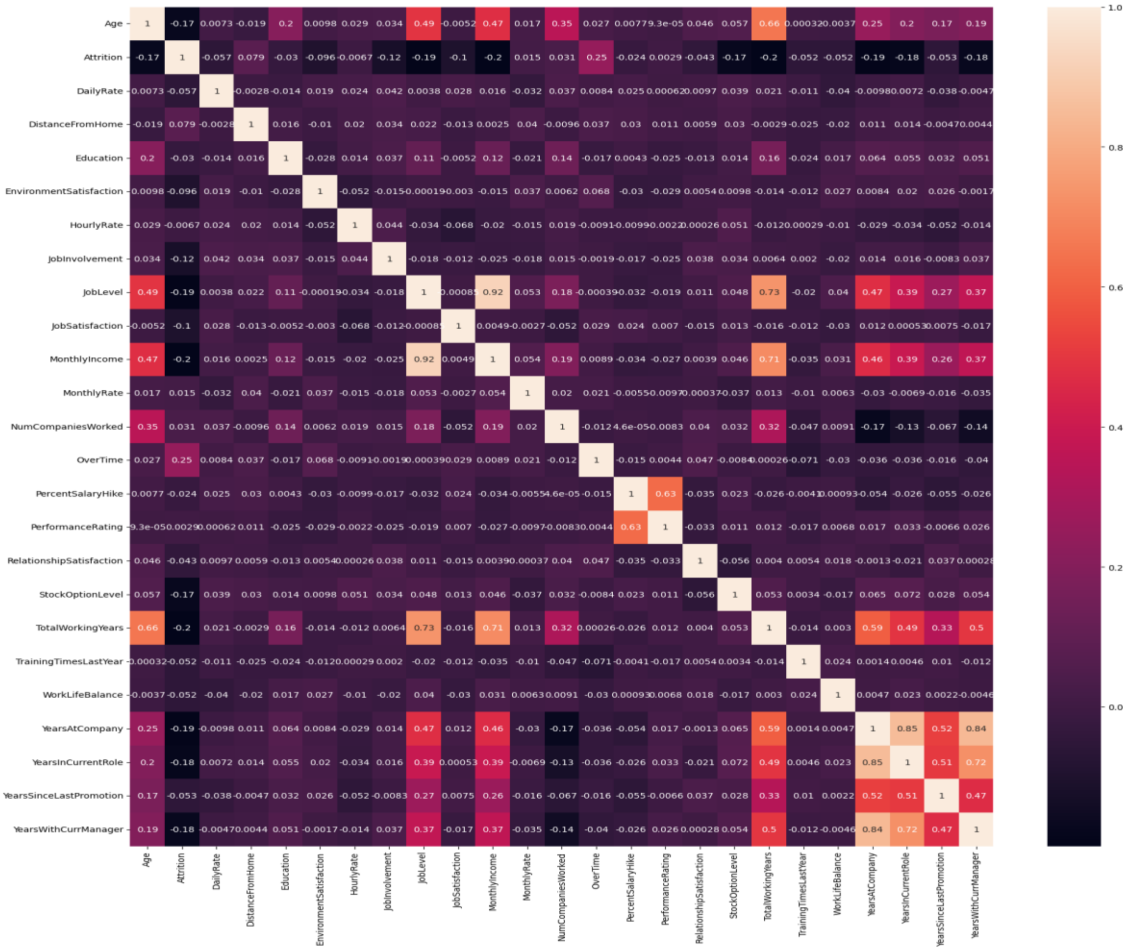
5.1. Comparison between those who stayed and left

Comparing the mean and standard deviation of the employees who stayed and left



- Age: Mean age of the employees who stayed is higher compared to who left
- Daily Rate: Rate of employees who stayed is higher
- Distance From Home: Employees who stayed live closer to home
- Environment Satisfaction & Job Satisfaction: Employees who stayed are generally more satisfied with their jobs

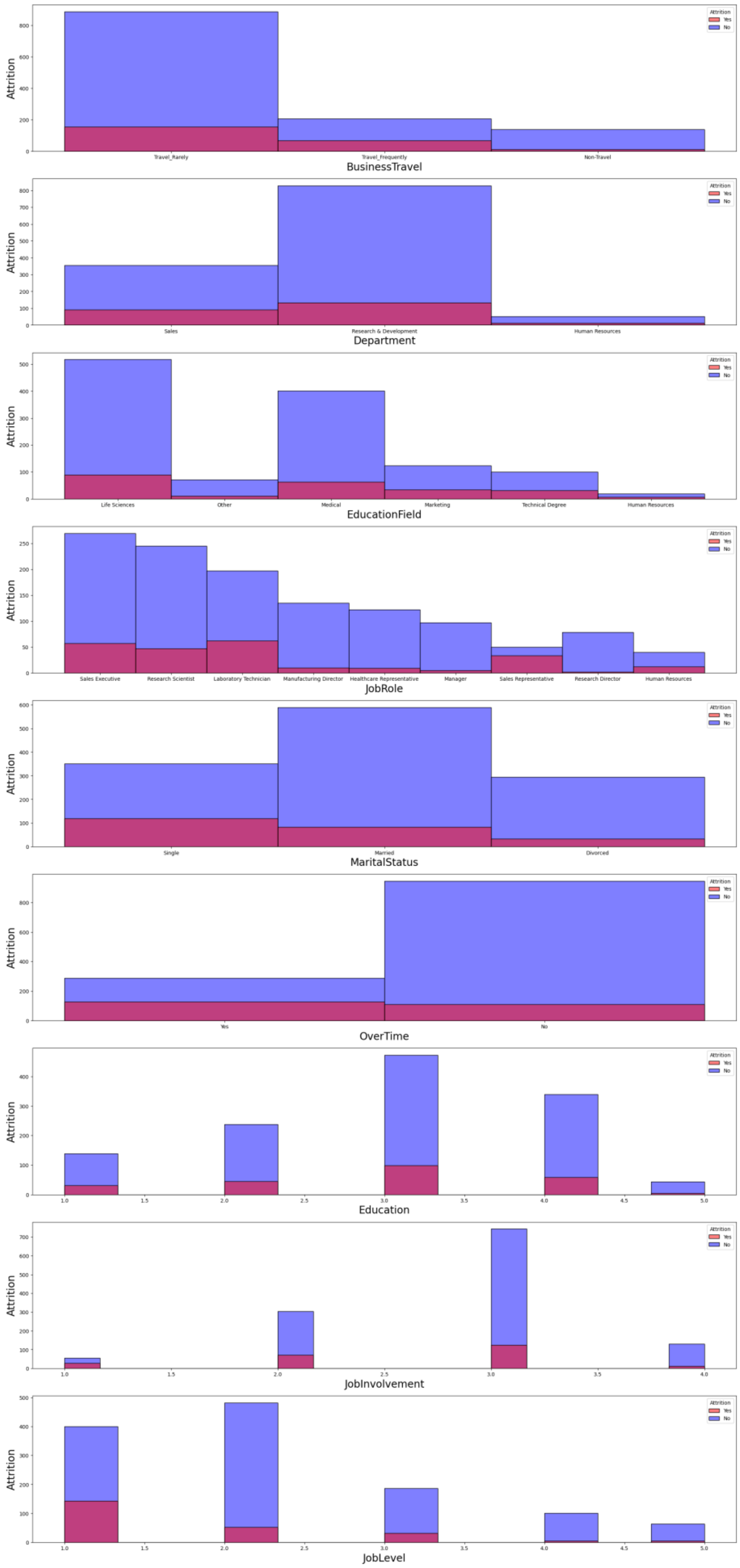
5.2. Corelation



- Job level is strongly correlated with total working hours
- Monthly income is strongly correlated with Job level
- Monthly income is strongly correlated with total working hours
- Age is strongly correlated with monthly income

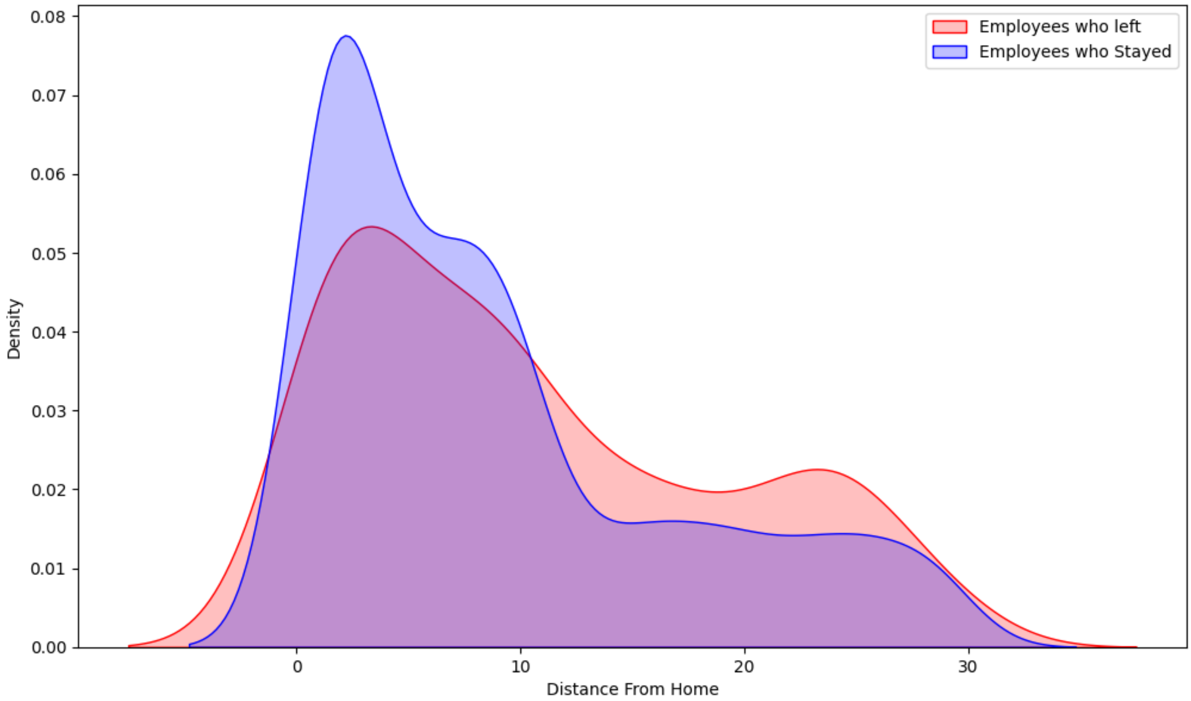
5.3. Who tend to leave

- Business Travel:-> Those who travel frequently they try to leave more
- Department:-> From Salse Department people try to leave more
- Education Field:-> Those who are in Marketing and Technical Degree they tend to leave more.
- Job Role:-> Sales Representative trying to leave more.
- Marital Status:-> Single employees tend to leave compared to married and divorced
- Over Time:-> Those who doing Over Time(Yes) they tend to leave more.
- Education:-> Bachelor's tend to leave more
- Job Involvement:-> Less involved employees tend to leave the company.
- Job Level:->Less experienced (low job level) tend to leave the company.


Attrition
Yes
No
Attrition
0
200
400
600
800
Travel_Rarely
Travel_Frequently
Non-Travel
BusinessTravel
0
100
200
300
400
500
600
700
800
Sales
Research & Development
Human Resources
Department
0
100
200
300
400
500
Life Sciences
Other
Medical
Marketing
Technical Degree
Human Resources
EducationField
0
50
100
150
200
250
Sales Executive
Research Scientist
Laboratory Technician
Manufacturing Director
Healthcare Representative
Manager
Sales Representative
Research Director
Human Resources
JobRole
0
100
200
300
400
500
600
Single
Married
Divorced
MaritalStatus
0
200
400
600
800
Yes
No
OverTime
0
100
200
300
400
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
Education
0
100
200
300
400
500
600
700
1.0
1.5
2.0
2.5
3.0
3.5
4.0
JobInvolvement
0
100
200
300
400
500
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
JobLevel

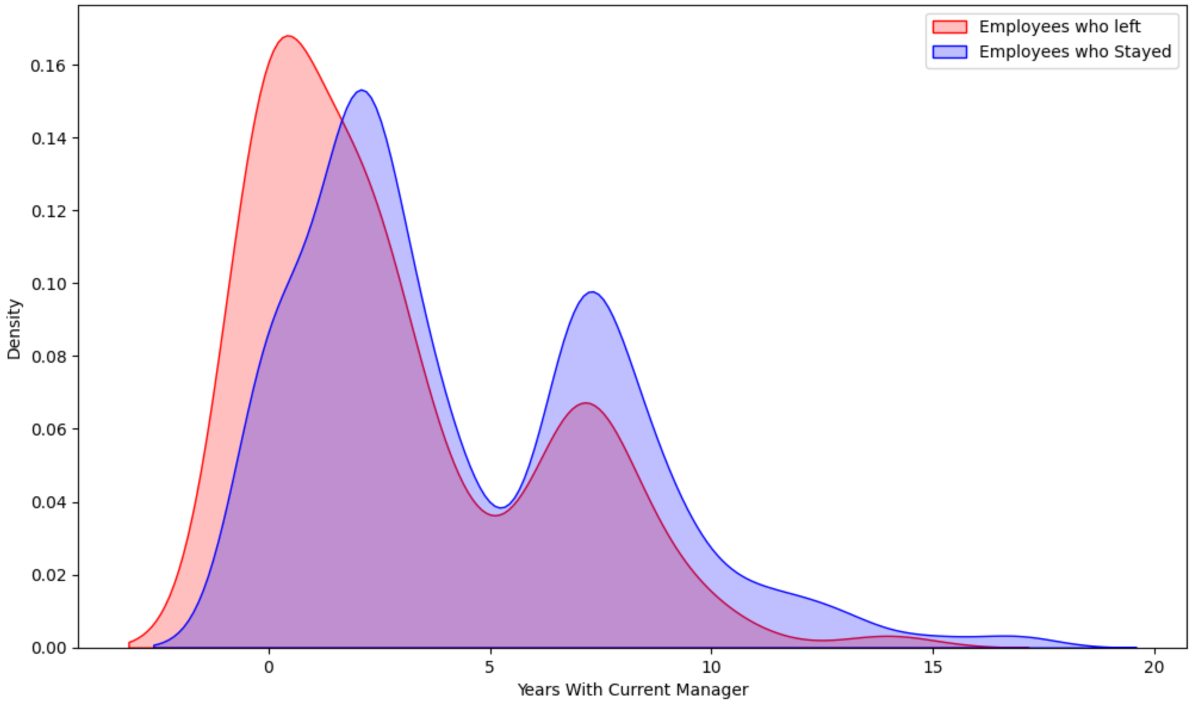
5.4. Other Significant things

- There is significant difference in the distance from home between employees who left and stayed (p<.05)
- Mann-Whitney's test to check if there is a significant difference between the two groups
- p-value is 0.0023870470273627984 which is less than 0.05, so we reject the null hypothesis
- There is significant difference in the distance from home between employees who left and stayed

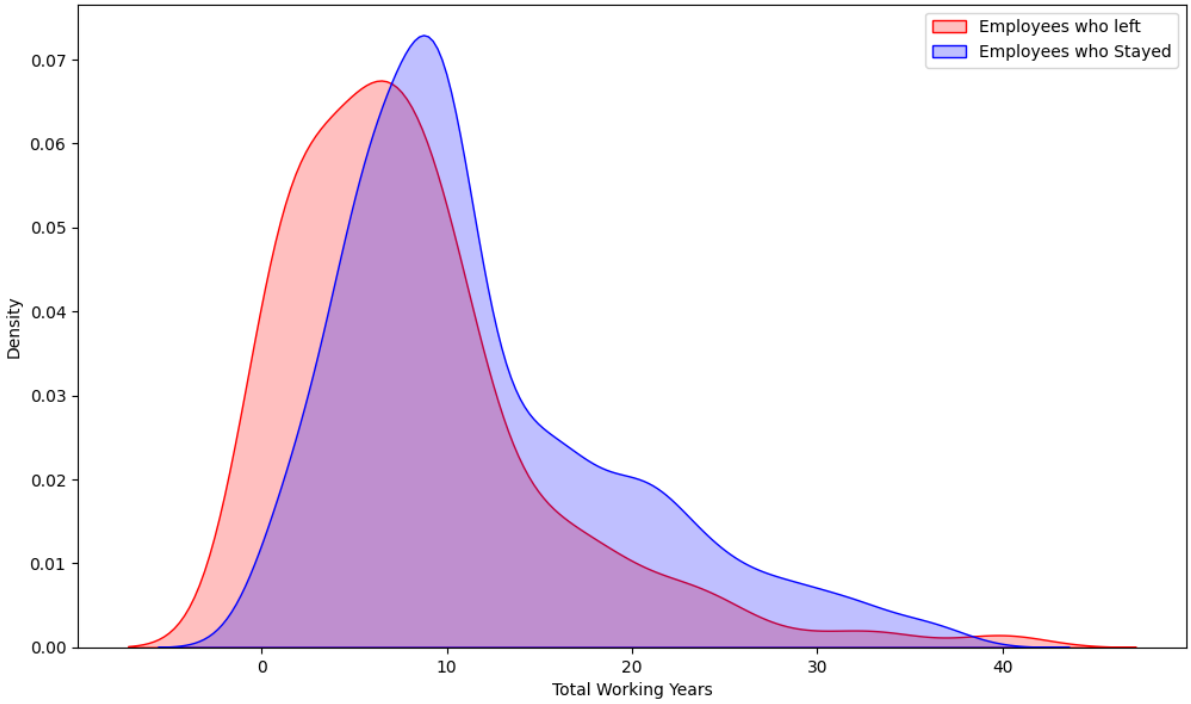


- There is significant difference in the Years With Current Manager between employees who left and stayed (p<.05)
- p-value is 1.8067542583144407e-11 which is less than 0.05, so we reject the null hypothesis

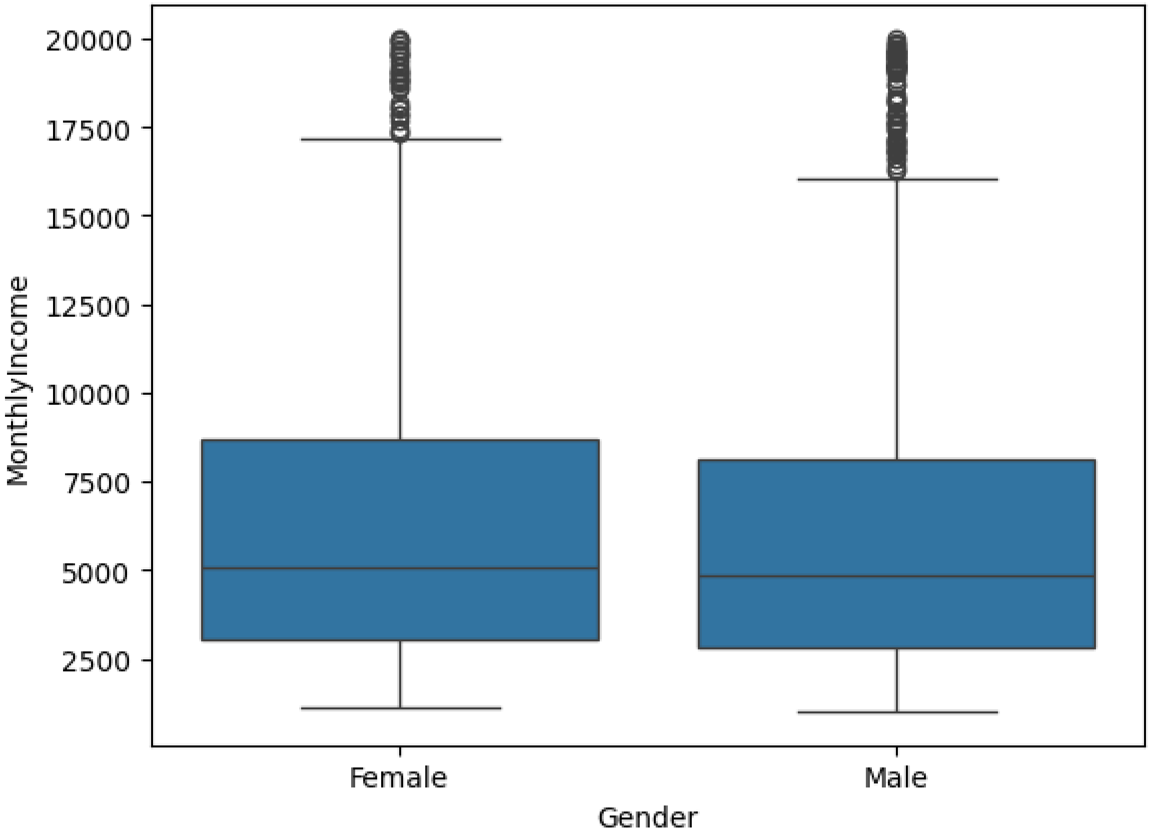
- There is significant difference in the years with current manager between employees who left and stayed



- There is significant difference in the Total Working Years between employees who left and stayed (p<.05)
- p-value is 2.399569364798952e-14 which is less than 0.05, so we reject the null hypothesis
- There is significant difference in the total working years between employees who left and stayed

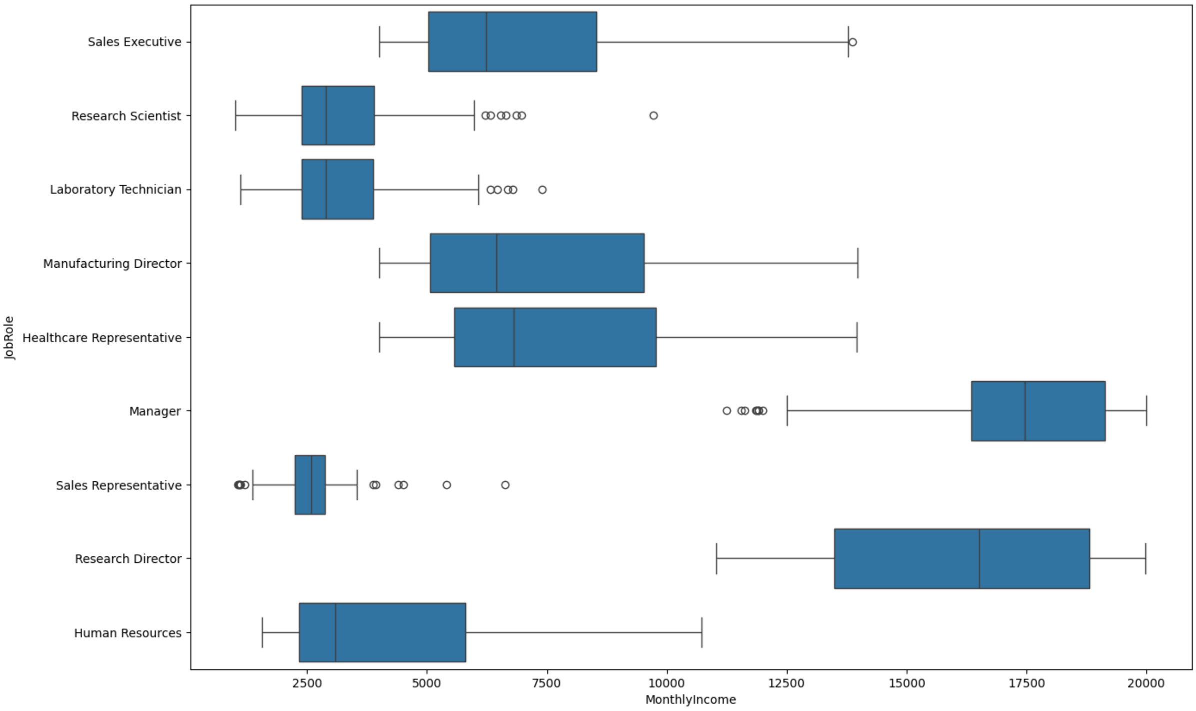


- There are no significant differences in Monthly Income between Female and Male employees (p=0.09)
- p-value is 0.08841668326602112 which is greater than 0.05, so we fail to reject the null hypothesis and assume no differences in MonthlyIncome between Male and Female employees



- Research Directors and Managers have the highest Monthly Income

- Sales Representatives have the lowest Monthly Income, followed by Research Scientists and Lab Technicians



6. Data Preparation

6.1. Dropping Columns

- 'EmployeeCount', 'StandardHours', 'Over18',and 'EmployeeNumber' column are drop first because there value is same for all employees except 'EmployeeNumber' but it's is not so relevant.

6.2. One Hot Encoding

- One hot Encoding is done in categorical which are 'BusinessTravel', 'Department', 'EducationField', 'Gender', 'JobRole',and 'MaritalStatus'.
- Then concatenated with numerical columns.

6.3. Data Scaling

- Min Max Scaler is used in scaling process.

6.4. Train and Test datasets

- Data is divided into train test using function 'train_test_split()', test size is used 25% for this dataset.

6.5. Class Balancing

- SMOTE is used in class balancing.
- Then new train test dataset is created which is copy of old train test but this data is converted in tensor so that it can use later for Simple Nural Network.

7. Model Selection and Training

7.1. Model Selection

- 'Logistic Regression', 'SVC', 'Random Forest Classifier', 'Decision Tree Classifier', 'XGB Classifier', 'CatBoost Classifier', 'K Neighbors Classifier' ,'Gaussian NB' ,'Gradient Boosting Classifier' ,'AdaBoost Classifier' ,and 'Simple Nural Network Binary Classification model' models are taken in selecting process for selecting the best model for this dataset.
- 'Simple Nural Network Binary Classification model' is made with the help of PyTorch library.

  (1) Logistic Regression

  (a) Accuracy score of Logistic Regression is: 0.829004329004329 and 0.7635869565217391
  (b) F1 score of Logistic Regression is: 0.47904191616766467
  (c) Precision score of Logistic Regression is: 0.37037037037037035
  (d) Recall score of Logistic Regression is: 0.6779661016949152

  (2) SVC

  (a) Accuracy score of SVC is: 0.954004329004329 and 0.8315217391304348
  (b) F1 score of SVC is: 0.515625
  (c) Precision score of SVC is: 0.4782608695652174
  (d) Recall score of SVC is: 0.559322033898305

  (3) Random Forest Classifier

  (a) Accuracy score of Random Forest Classifier is: 1.0 and 0.8559782608695652
  (b) F1 score of Random Forest Classifier is: 0.4044943820224719
  (c) Precision score of Random Forest Classifier is: 0.6
  (d) Recall score of Random Forest Classifier is: 0.3050847457627119

  (4) Decision Tree Classifier

  (a) Accuracy score of Decision Tree Classifier is: 1.0 and 0.7744565217391305
  (b) F1 score of Decision Tree Classifier is: 0.366412213740458
  (c) Precision score of Decision Tree Classifier is: 0.3333333333333333
  (d) Recall score of Decision Tree Classifier is: 0.406779661016949

  (5) XGB Classifier

  (a) Accuracy score of XGB Classifier is: 1.0 and 0.8804347826086957
  (b) F1 score of XGB Classifier is: 0.5686274509803921
  (c) Precision score of XGB Classifier is: 0.6744186046511628
  (d) Recall score of XGB Classifier is: 0.4915254237288136

  (6) Cat Boost Classifier

  (a) Accuracy score of Cat Boost Classifier is: 1.0 and 0.875
  (b) F1 score of Cat Boost Classifier is: 0.5306122448979592
  (c) Precision score of Cat Boost Classifier is: 0.6666666666666666

(d) Recall score of Cat Boost Classifier is: 0.4406779661016949

(7) K Neighbors Classifier

(a) Accuracy score of K Neighbors Classifier is: 0.8804112554112554 and 0.6630434782608695

(b) F1 score of K Neighbors Classifier is: 0.38613861386138615

(c) Precision score of K Neighbors Classifier is: 0.2727272727272727

(d) Recall score of K Neighbors Classifier is: 0.6610169491525424

(8) Gaussian NB

(a) Accuracy score of Gaussian NB is: 0.7083333333333334 and 0.5597826086956522

(b) F1 score of Gaussian NB is: 0.31932773109243695

(c) Precision score of Gaussian NB is: 0.2122905027932961

(d) Recall score of Gaussian NB is: 0.6440677966101694

(9) Gradient Boosting Classifier

(a) Accuracy score of Gradient Boosting Classifier is: 0.9788961038961039 and 0.8614130434782609

(b) F1 score of Gradient Boosting Classifier is: 0.49504950495049505

(c) Precision score of Gradient Boosting Classifier is: 0.5952380952380952

(d) Recall score of Gradient Boosting Classifier is: 0.423728813559322

(10) Ada Boost Classifier

(a) Accuracy score of Ada Boost Classifier is: 0.9296536796536796 and 0.8505434782608695

(b) F1 score of Ada Boost Classifier is: 0.5528455284552846

(c) Precision score of Ada Boost Classifier is: 0.53125

(d) Recall score of Ada Boost Classifier is: 0.576271186440678

(11) Simple Nural Network

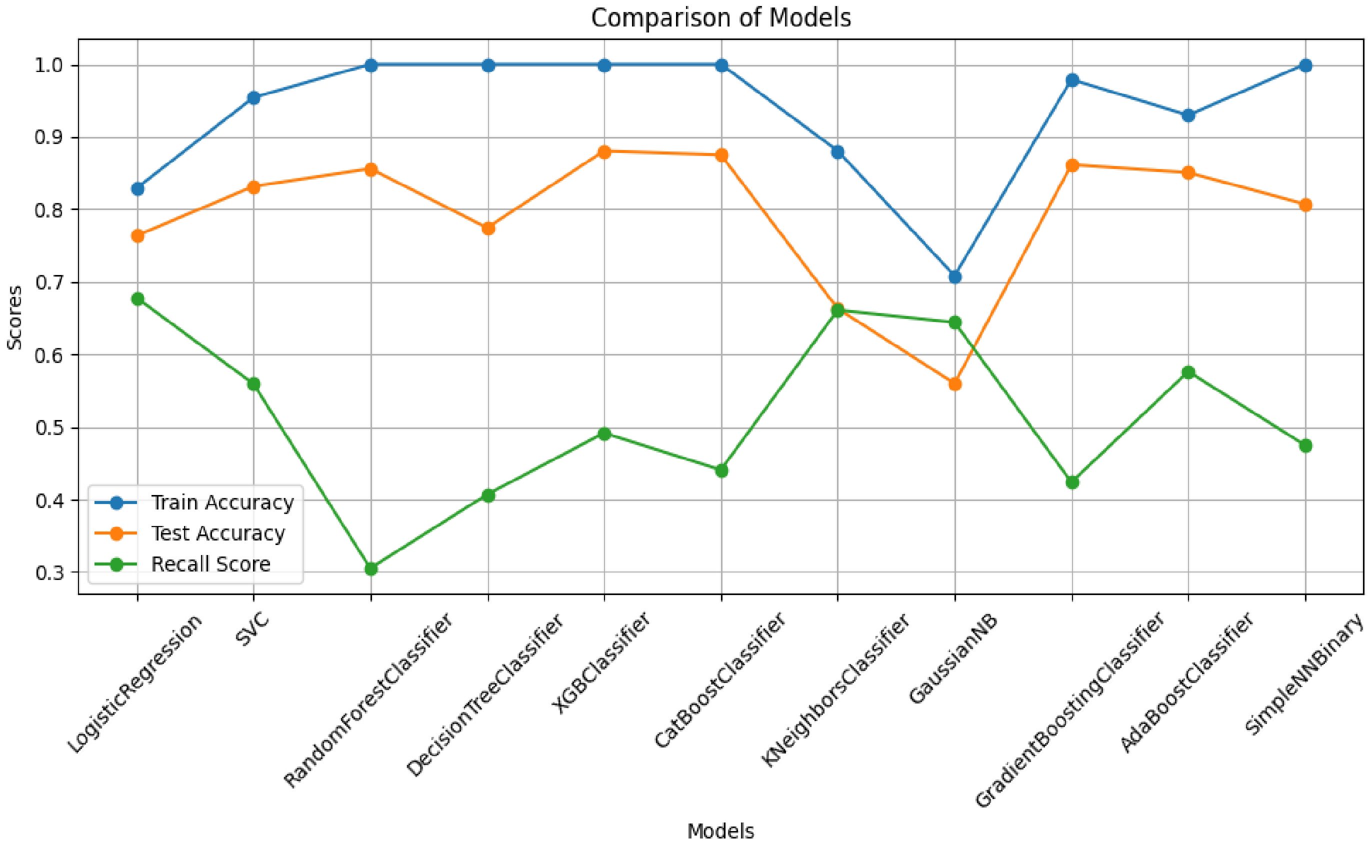
(a) Accuracy on training set: 100.00%

(b) Accuracy on test set: 80.71%

(c) Recall on test set: 47.46%

(d) F1 score is: 0.4409448818897638

(e) Precision score is: 0.4117647058823529



8. Hyperparameter Tuning

- With the help of Grid Search these parameters are finded for Logistic Regression
- 'C': 1
- 'class_weight': None
- 'fit_intercept': True
- 'max_iter': 100
- 'penalty': 'l1'
- 'solver': 'liblinear'
- 'tol': 0.0001
- Logistic Regression values become:-> train accuracy 80%, test accuracy 80%, F1 score 0.48 and recall value 0.68

9. Conclusion:

LR showed a much better recall for the minority class (0.68). Even with the SMOTE balancing, models showed a low True Positive Rate for the minority class. This is most likely due to imbalanced clases and lack of data to train the model, but because it depends on the number of company's employees, we will have to stick with LR model as it showed the best results of all. On the other hand, in this case we give more importance to the correct prediction of employee atrittion and we do not care much about false positives (employees predicted as leaving the company but actually stay).

10. References:

1. www.kaggle.com
2. scikit-learn.org
3. https://www.ibm.com/in-en