Human Resources
Employee Churn Prediction
Report

Group 1

JAHNAVI REDDY GANESINA

ganesina.j@northeastern.edu (+1 5134179172)

Percentage of Effort Contributed by Student 1: 100%

Signature of Student 1: J.R.G

Submission Data: March 21, 2024

Table of Contents

Project Setting

In today's business environment, managing employee turnover has become a critical challenge for organizations, particularly within the realm of human resources management. High turnover rates not only result in increased recruitment and training costs but also impact organizational productivity and morale. The direct financial implications, such as severance pay and the investment in training new employees, are significant. However, the indirect effects, such as the erosion of institutional knowledge, disruption of team dynamics, and potential damage to the organization's reputation in the job market, can be even more detrimental.

This project seeks to thoroughly explore the phenomenon of employee churn within a specific industry or domain. By identifying the root causes of turnover, which may range from dissatisfaction with job roles, misalignment with organizational culture, to inadequate growth opportunities and compensation, the aim is to uncover targeted insights. These insights will inform the development of effective retention strategies tailored to the unique challenges and opportunities within the industry.

Effective retention strategies may include initiatives aimed at enhancing employee engagement, cultivating a positive and inclusive organizational culture, offering competitive compensation and benefits packages, and providing clear career advancement opportunities. The goal is to not only address the immediate issue of turnover but to foster an environment where employees are engaged, satisfied, and motivated to contribute to the organization's success over the long term.

By committing to a strategic approach to understanding and mitigating employee turnover, organizations can significantly improve retention rates. This not only benefits the organization in terms of reduced costs and improved productivity but also contributes to a more stable and positive work environment. Ultimately, this project aims to equip industry leaders with the knowledge and tools necessary to build a resilient, high-performing workforce that supports the organization's goals and fosters sustainable success.

Problem Definition

Organizations across various industries face the challenge of employee churn, which can lead to significant operational disruptions and increased costs related to recruitment, training, and loss of institutional knowledge. The specific problem addressed in this project is to not only understand and analyze the patterns of employee departure within the organization but also to develop a predictive model that can identify potential future churn. This insight is crucial for developing effective retention strategies and ensuring organizational stability and growth.

- Understanding Factors Contributing to Employee Churn: The first objective is to identify and understand the various factors that lead to employee departure. This involves analyzing employee data to pinpoint specific trends, behaviors, and circumstances that commonly precede an employee's decision to leave. Factors may include but are not limited to job satisfaction, career progression opportunities, work-life balance, management practices, and organizational culture.

- Predicting Future Churn: Leveraging the insights gained from analyzing departure patterns, the project aims to build a predictive model that can forecast which employees are at a higher risk of leaving. This model will employ statistical and machine learning techniques to evaluate employee data and predict churn likelihood based on identified patterns and factors.

- Developing Mitigation Strategies: With a predictive model in place, the project's focus will shift towards utilizing the insights to implement data-driven strategies aimed at reducing turnover rates. This could involve addressing identified issues contributing to churn, enhancing employee engagement and satisfaction, and creating a more supportive and fulfilling work environment.

Data Sources

Kaggle : https://www.kaggle.com/code/shifanaaz125/hr-dataset-eda/input

Data Description

Dataset contains 1471 records available with 35 columns. Below displayed are 17 out of the 35 columns in the dataset:

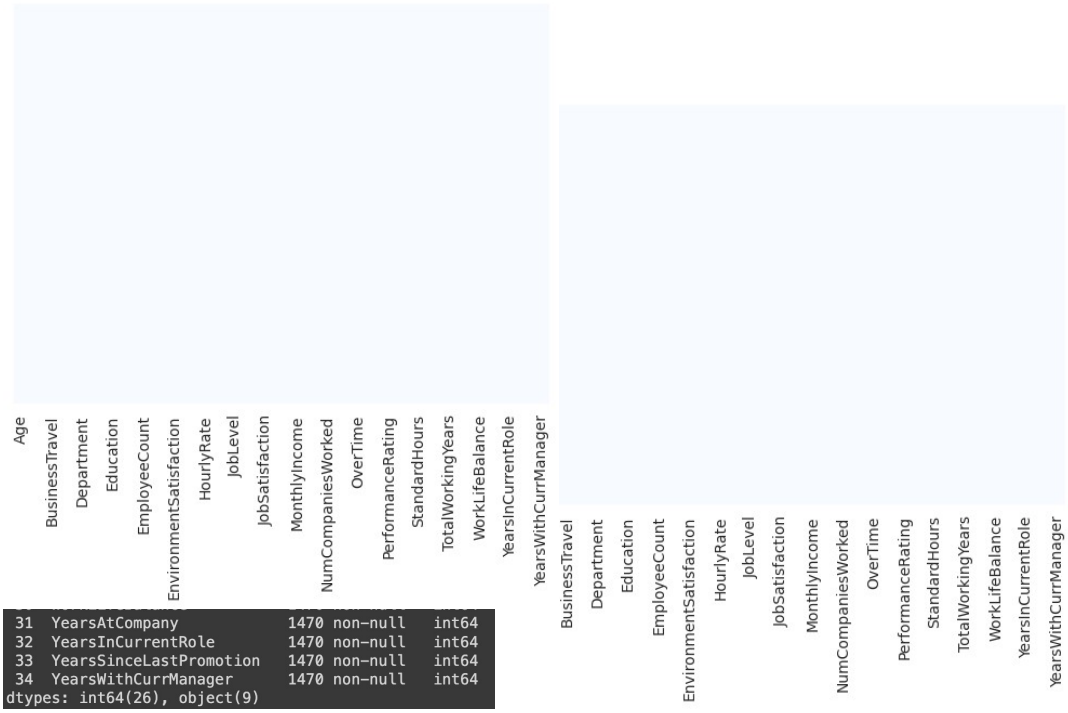
Age	Age in years of the employee [Int]
Attrition	People who people leave [Boolean]
BusinessTravel	Work related travel [Categorical]
DailyRate	Daily rate that employee is paid [Numerical]
Department	Department [Categorical]
DistanceFromHome	Distance from home to work [Numerical]
Education	Level of education of the employee [Categorical]
HourlyRate	Hourly rate of pay of the employee [Numerical]
JobInvolvement	Employee job involvement ratings [Numerical]
JobLevel	Employee Job level [Numerical]
JobRole	Employee Job role [Categorical]
JobSatisfaction	Employee Job Satisfaction [Numerical]
PercentSalaryHike	Salary increments in Percentages[Numerical]
PerformanceRating	Performance rating[Numerical]
RelationshipSatisfaction	Relationship satisfaction[Numerical]
StandardHours	Employee standard hours worked[Numerical]

Data Pre-processing:

1. Shape of the dataset: Below is the count of number of rows and columns.

```
Number of rows: 1470
Number of columns: 45
```

2. Dataset has no missing values.




```
 31  YearsAtCompany             1470 non-null   int64
 32  YearsInCurrentRole         1470 non-null   int64
 33  YearsSinceLastPromotion    1470 non-null   int64
 34  YearsWithCurrManager       1470 non-null   int64
dtypes: int64(26), object(9)
```

3. Below are the statistics of the variables in the dataset:

	Age	DailyRate	DistanceFromHome	Education	EmployeeCount	EmployeeNumber	EnvironmentSatisfaction	HourlyRate	JobInvolvement
count	1470.000000	1470.000000	1470.000000	1470.000000	1470.0	1470.000000	1470.000000	1470.000000	1470.000000
mean	36.923810	802.485714	9.192517	2.912925	1.0	1024.865306	2.721769	65.891156	2.729932
std	9.135373	403.509100	8.106864	1.024165	0.0	602.024335	1.093082	20.329428	0.711561
min	18.000000	102.000000	1.000000	1.000000	1.0	1.000000	1.000000	30.000000	1.000000
25%	30.000000	465.000000	2.000000	2.000000	1.0	491.250000	2.000000	48.000000	2.000000
50%	36.000000	802.000000	7.000000	3.000000	1.0	1020.500000	3.000000	66.000000	3.000000
75%	43.000000	1157.000000	14.000000	4.000000	1.0	1555.750000	4.000000	83.750000	3.000000
max	60.000000	1499.000000	29.000000	5.000000	1.0	2068.000000	4.000000	100.000000	4.000000

4. Number of unique values for each Categorical columns.

```
Attrition          2
BusinessTravel     3
Department         3
EducationField     6
Gender             2
JobRole            9
MaritalStatus      3
Over18             1
OverTime           2
dtype: int64
```

5. Changed Attrition values from ‘Yes’ to 1 and ‘No’

Data Exploration

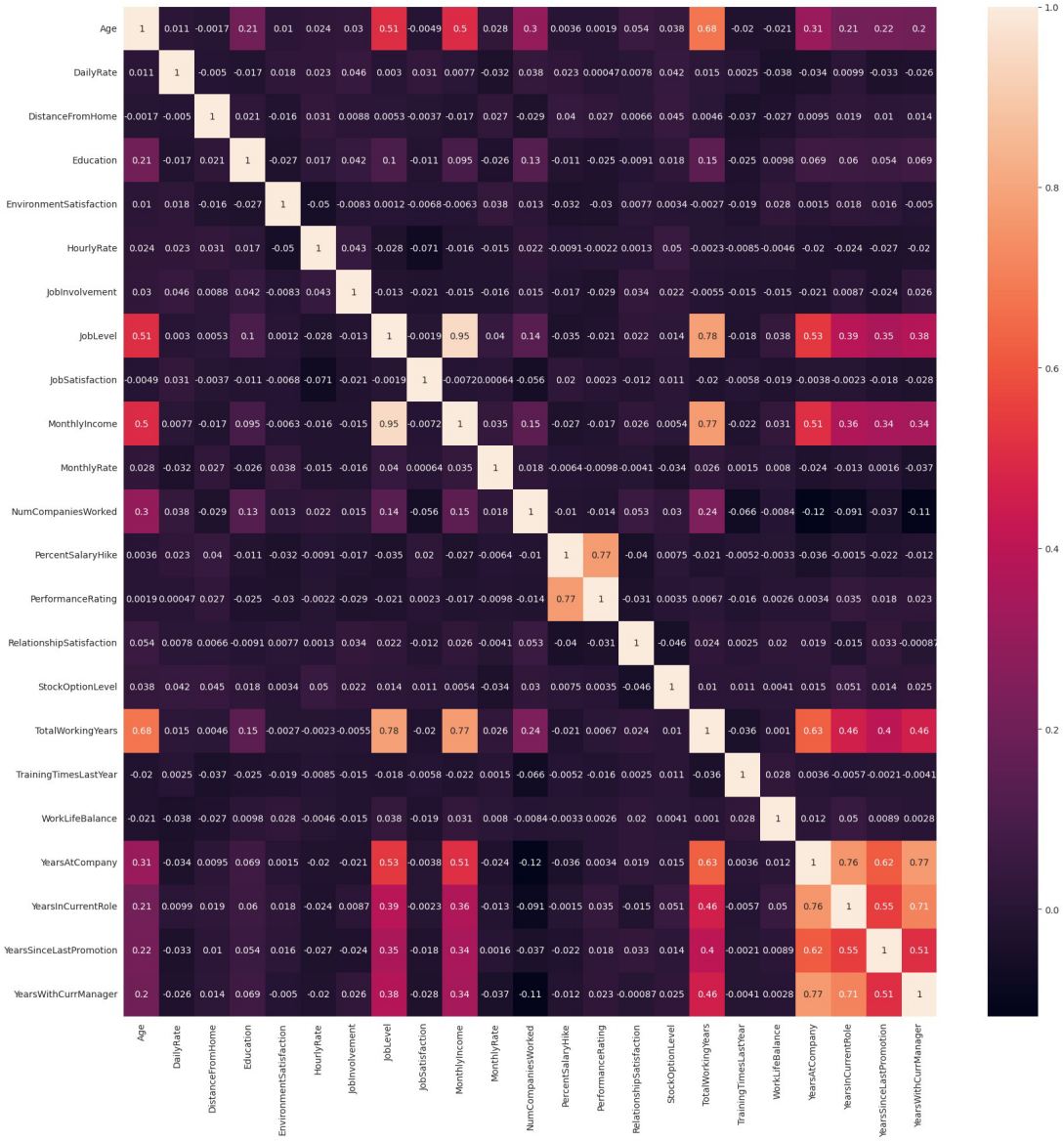
1. Histogram

We created histograms for each numerical variable in the DataFrame. The histograms help us understand the spread and central tendency of each variable.

2. Dimension Reduction

Dropping columns 'EmployeeCount', 'StandardHours', 'Over18', 'EmployeeNumber' because they do not change from one employee to another.

3. Correlation Matrix



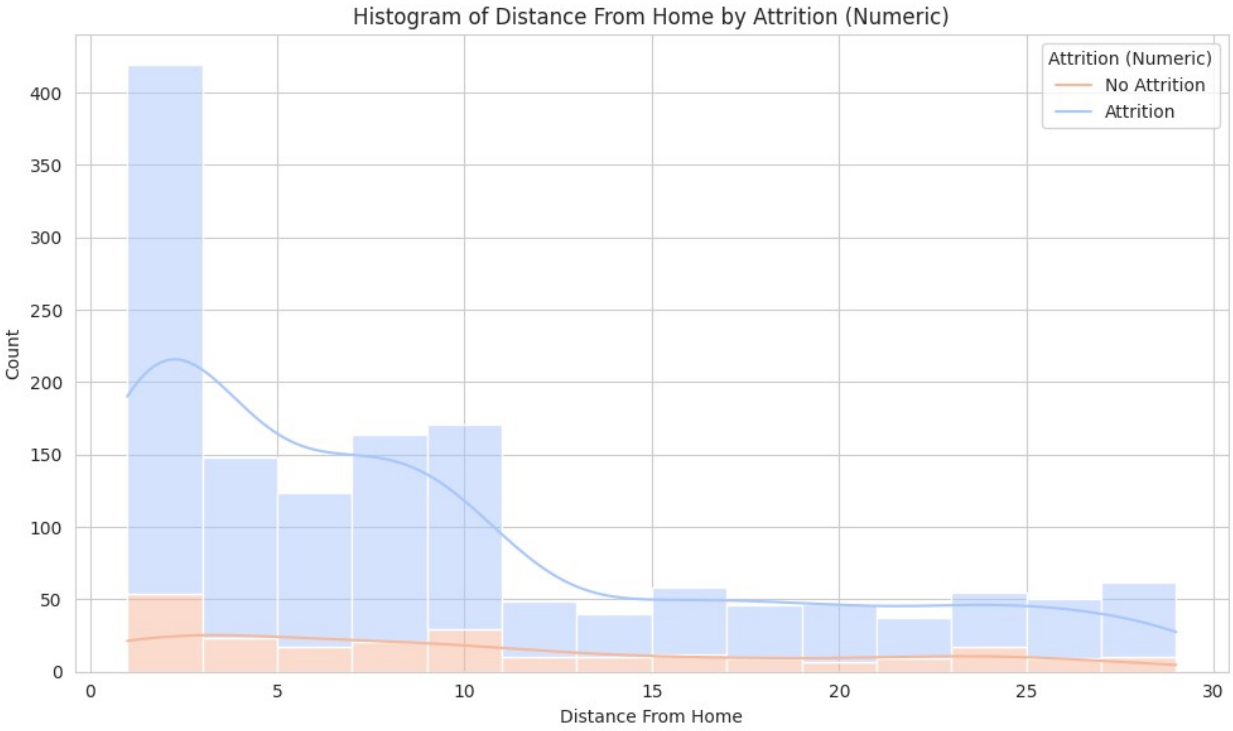
- The level of one's job exhibits a robust correlation with the overall number of working hours expended, indicating a clear association between job hierarchy and the extent of time devoted to work-related activities.
- There exists a strong correlation between the monthly income an individual earns and their job level, underscoring the significance of hierarchical position in influencing income. Moreover, the amount of monthly income demonstrates a

pronounced correlation with the total working hours invested, emphasizing the impact of time commitment on financial remuneration.

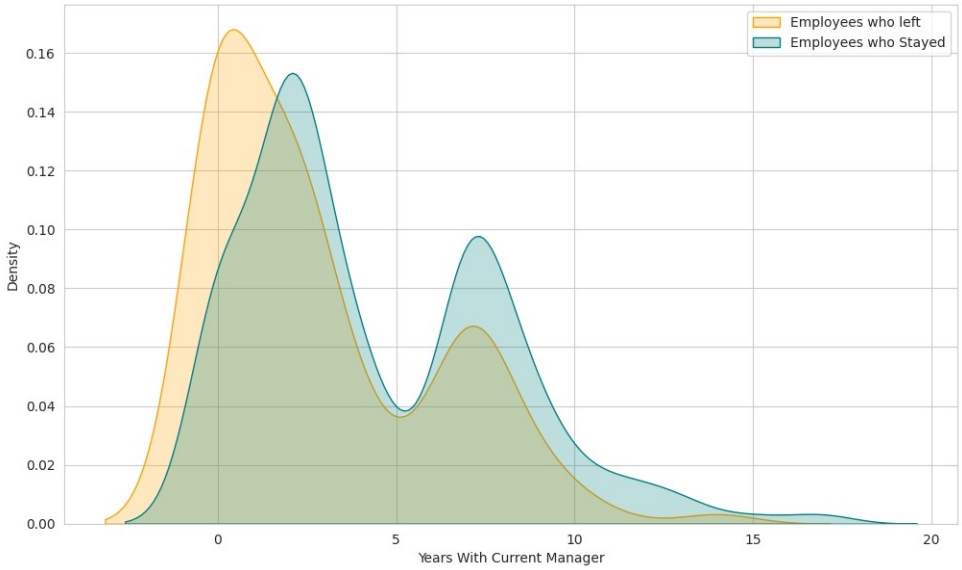
- Age exhibits a substantial correlation with monthly income, indicating a discernible link between an individual's age and their earning capacity.

4. Histogram with KDE

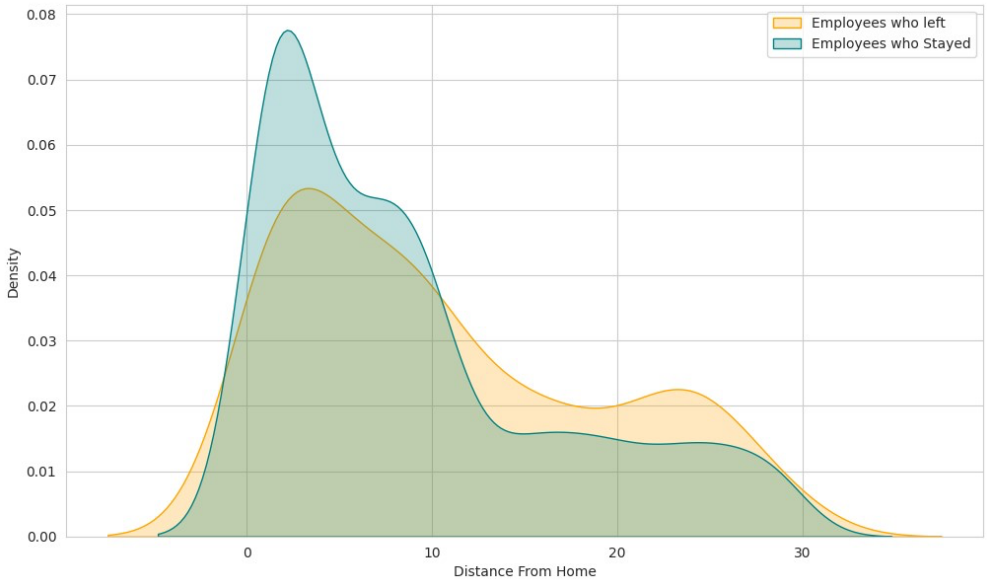
In terms of distance distribution, most employees tend to reside within a 5-mile radius of their workplace, indicating a common preference for proximity. Surprisingly, there is a higher attrition rate among those in closer proximity, challenging the assumption that longer commutes lead to more attrition. As distance increases beyond 10 miles, attrition rates decline.



5. KDE



- Significant peak is observed at 2-3 years, indicating a critical attrition period with current managers.
- Conversely, employees who stay peak around 4-5 years, with a secondary peak at 7 years, suggesting prolonged longevity.

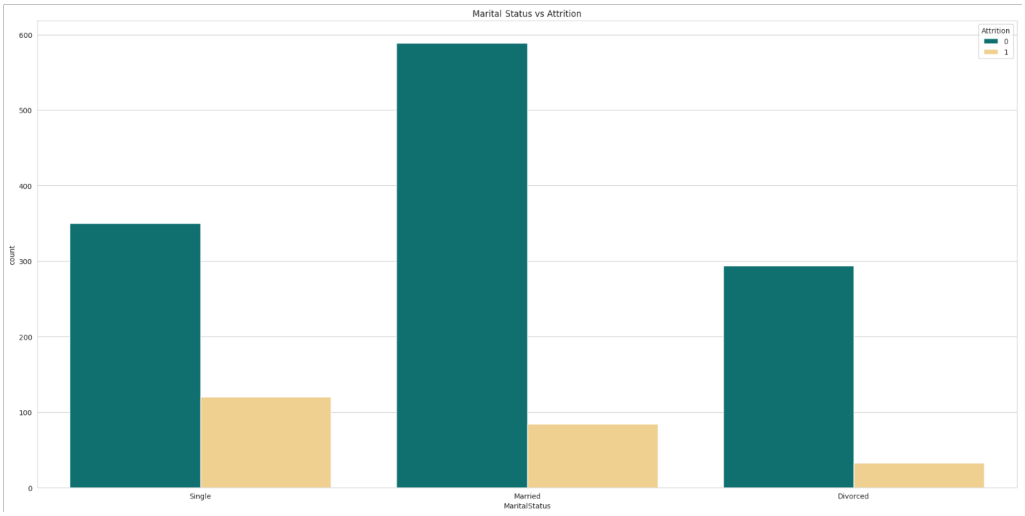


- The primary commuting distance for both departing and remaining employee's peaks at around 10 miles, indicating it may not be the sole factor influencing their decisions.

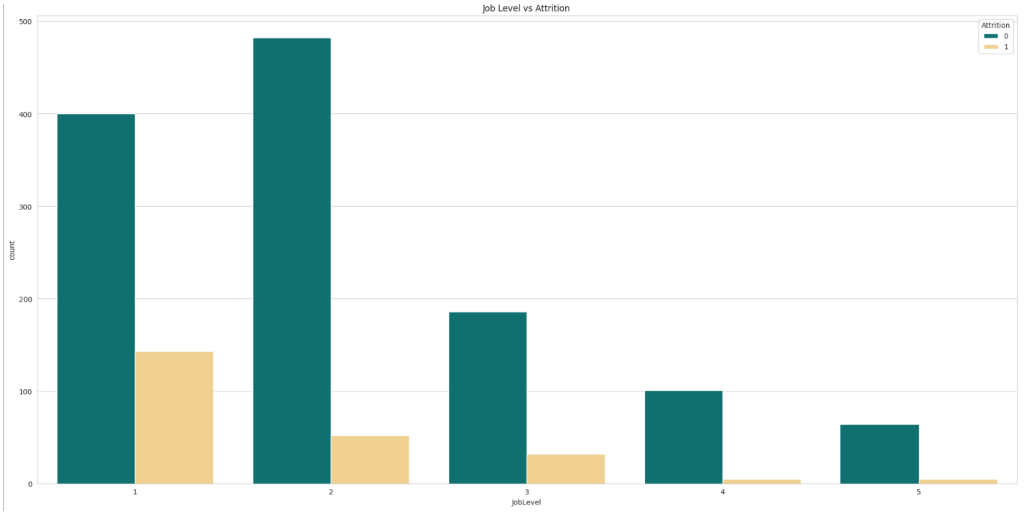
- Departing employees show a concentration in shorter distances and a notable attrition peak for those living very close to work.

6. Count Plots

The first plot shows the distribution of attrition across different marital statuses, while the second plot shows the distribution across job levels.

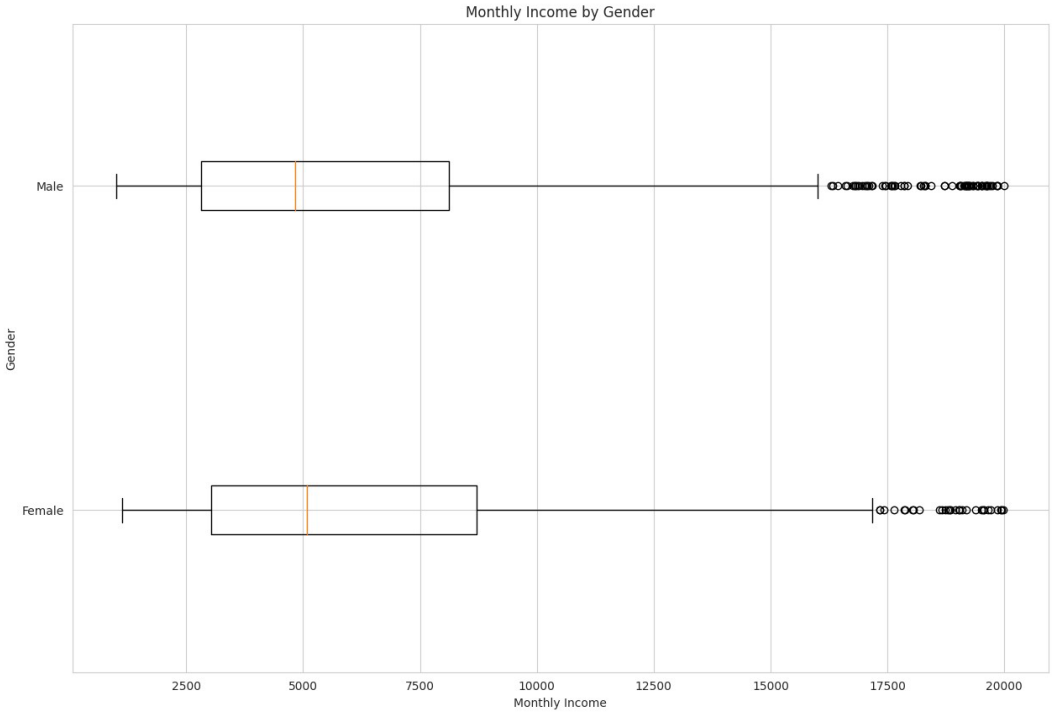


- Single employees show a notably higher attrition rate than married and divorced counterparts.



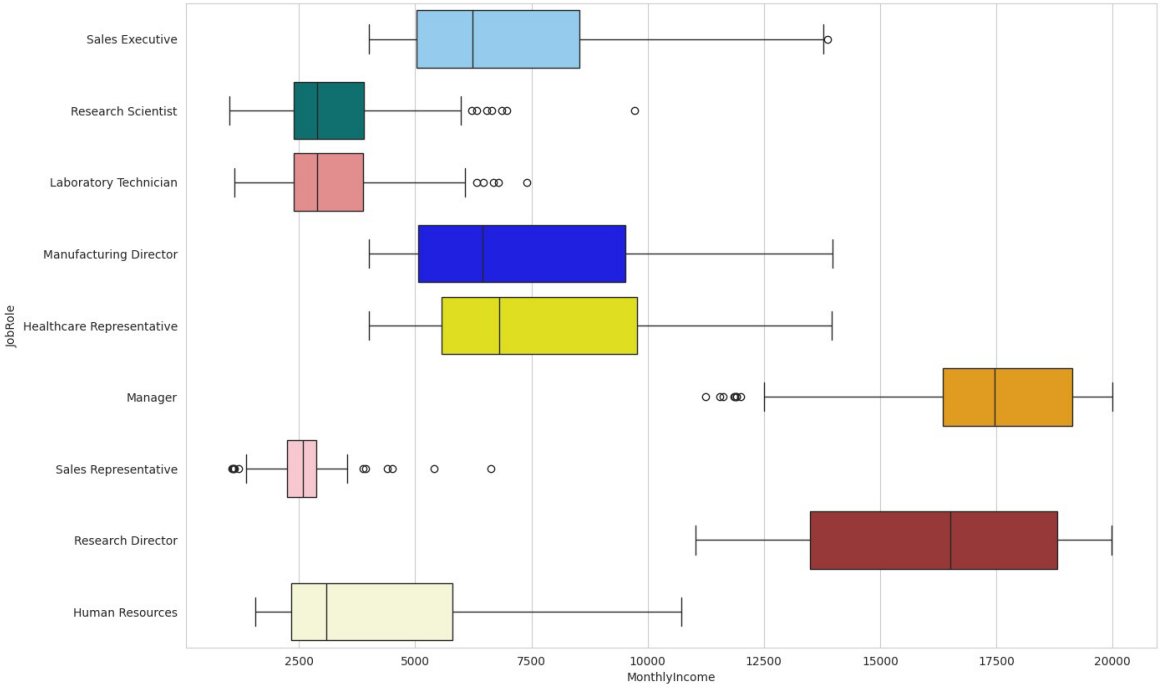
- Employee in lower job level tend to leave or switch companies more often.

7. Box Plot



- Males have a slightly higher median monthly income compared to females, as indicated by the median line within the box.
- There is a wider range of income levels among males, with more extreme values on the higher end, as shown by the longer upper whisker and more outliers above the upper quartile.

8. Boxplot



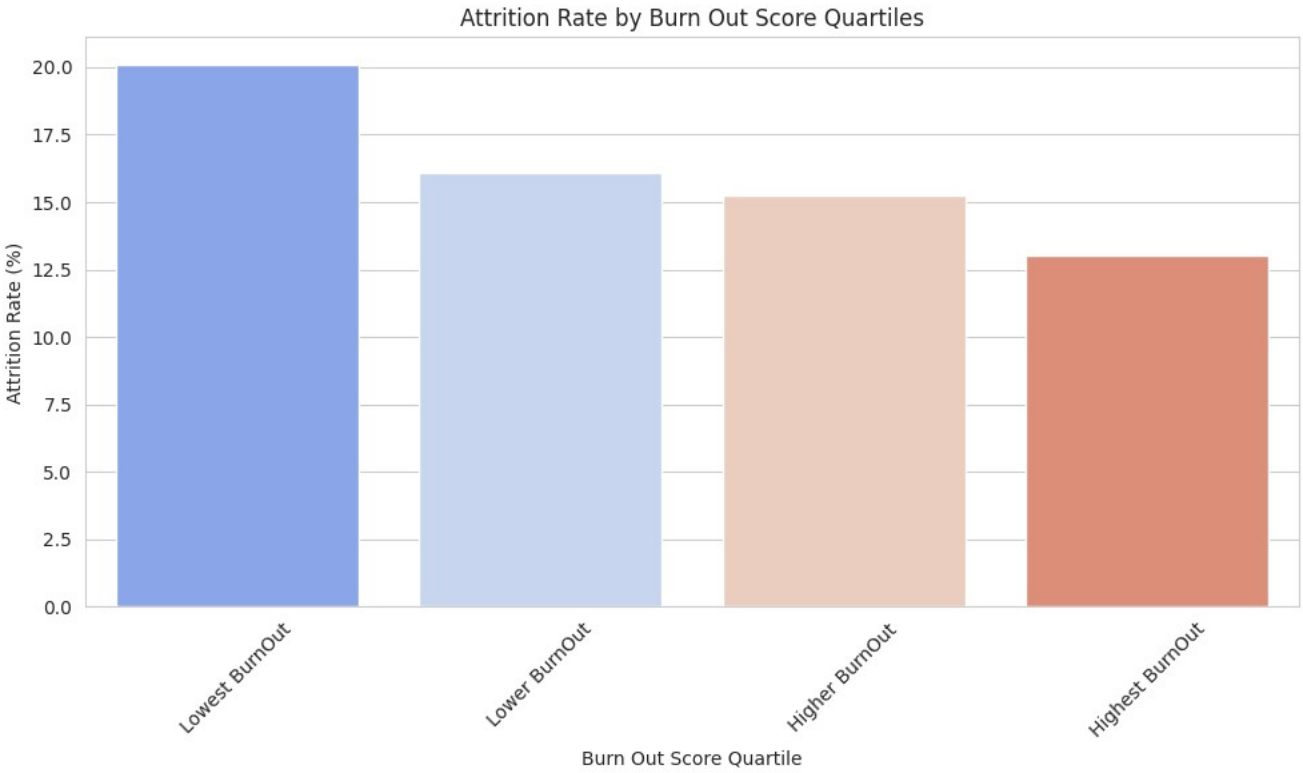
- Income Variation by Role: There's a clear variation in median monthly incomes across different job roles, with roles like 'Manager' and 'Research Director' having higher medians compared to others like 'Sales Representative' and 'Laboratory Technician'.

- Outliers: Several job roles have outliers indicating that there are employees who earn significantly more than their peers within the same job role, particularly noticeable in 'Sales Executive', 'Research Scientist', and 'Manager'.

- Income Range: The job roles of 'Manager', 'Manufacturing Director', and 'Healthcare Representative' show a wider range of incomes (as indicated by the length of their boxes and whiskers), suggesting greater diversity in compensation within these roles.

9. **Burnout Score Quartiles**

Burnout is a state of chronic stress leading to physical and emotional exhaustion, detachment, and a sense of ineffectiveness.



- Key insights from the analysis suggest that paradoxically, employees with higher "Burn Out" scores, potentially indicating more stress, have lower attrition rates. This unexpected trend suggests that burnout may not be the primary factor driving attrition, highlighting the need for a deeper exploration of underlying causes.

Data Mining Models

1) Logistic Regression Classifier:

Logistic regression is used to predict binary output. In this case, attrition prediction, whether the employee would leave the company or not considering various factors like DailyRate, BusinessTravel, DistanceFromHome etc.This model's algorithm works by implementing a linear equation first with independent predictors to predict the value. Then, the value will be converted to probability ranging between 0 to 1.

Data partition: Training (80%) and Testing (20%)

To evaluate the performance of this model, performance metrics like F1-score, accuracy, recall, precision can be calculated.

2) Random Forrest Classifier:

Decision Trees are supervised Machine Learning technique where data is split according to certain condition/ parameter. Set of decision trees are created from randomly selected subset of training set. Further, combined via voted from different decision trees to decide final class of test object. After cleaning the dataset, there are 30 input variables which will be the predictor variables, and target variable will the 'Attrition'.

Data partition: Training (80%) and Testing(20%)

3) Neural Networks:

The model will include 'adam' as optimizer, 'relu' and 'sigmoid' as activation functions. Sigmoid gives probabilities ranging between 0 to 1.

Evaluation metric used to for understanding performance will be 'Accuracy' and 'Binary crossentropy' (loss function) to measure the difference between predicted binary outcomes and actual binary value.

Data Partition: Training (70%), Validation(20%) and, Test(10%)

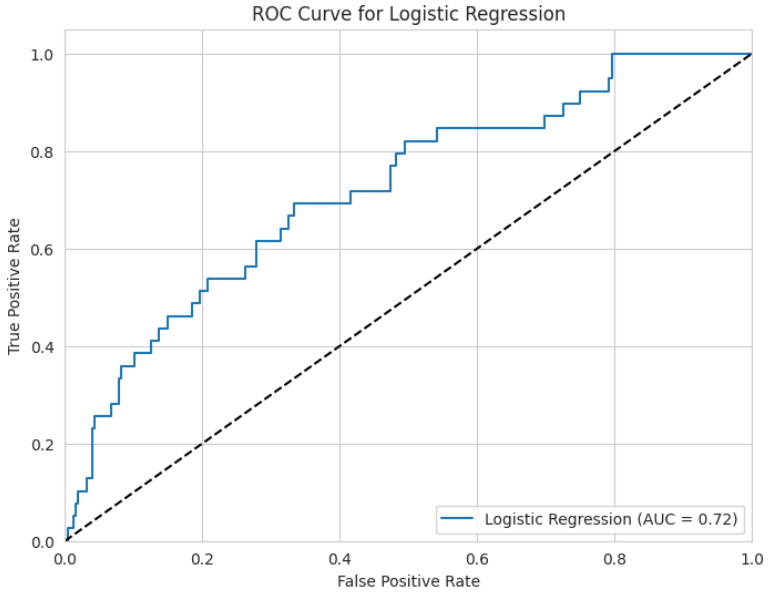
Model	Advantages	Disadvantages
Logistic Regression Classifier	(i) Good for linearly separable data. (ii) Can be updated easily with new data using stochastic gradient descent.	(i) Not suitable for complex relationships or non-linear data. (ii) Can be outperformed by more complex models on large datasets.
Random Forrest Classifier	(i) Handles non-linear data well. (ii) Can deal with missing values and maintains accuracy for missing data.	(i) More complex than logistic regression. (ii) Requires more memory and computational power.
Neural Networks	(i) Can work with unstructured data like images and text. (ii) Good at capturing interactions between features.	(i) Requires large amounts of data to perform well. (ii) Requires careful tuning of parameters and architecture.

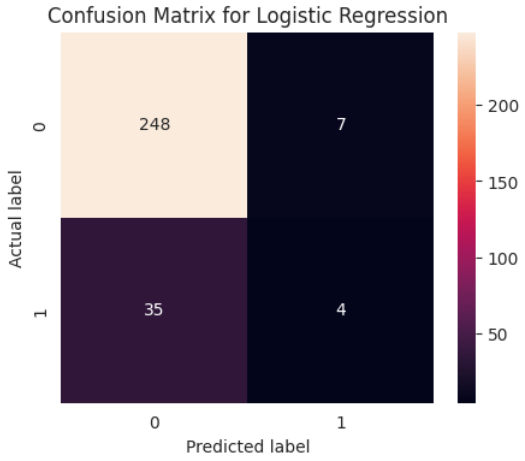
Model Performance Evaluation and Interpretation

1) Logistic Regression Classifier:

Data partition: Training (80%) and Testing (20%)

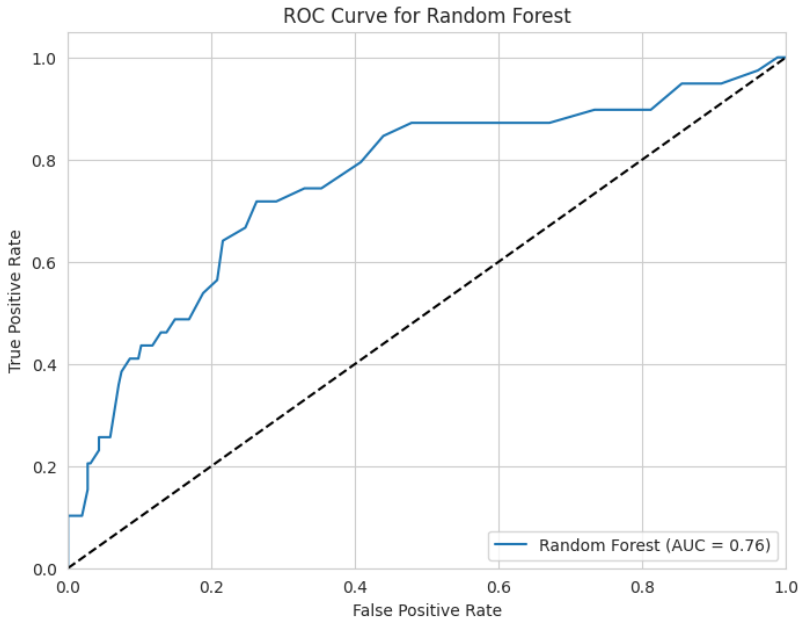
Accuracy of this model: 85.7% which is quite good.

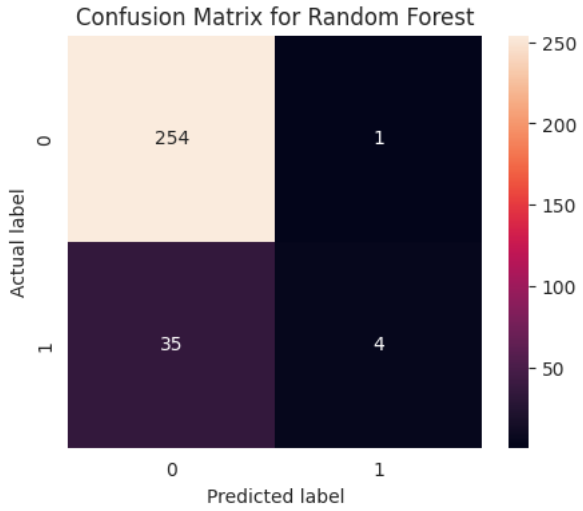




2) Random Forrest Classifier

Data partition: Training (80%) and Testing (20%). Similar partition as Logistic Regression. However, the accuracy of this model(87.7%) is slightly higher than Logistic Regression.



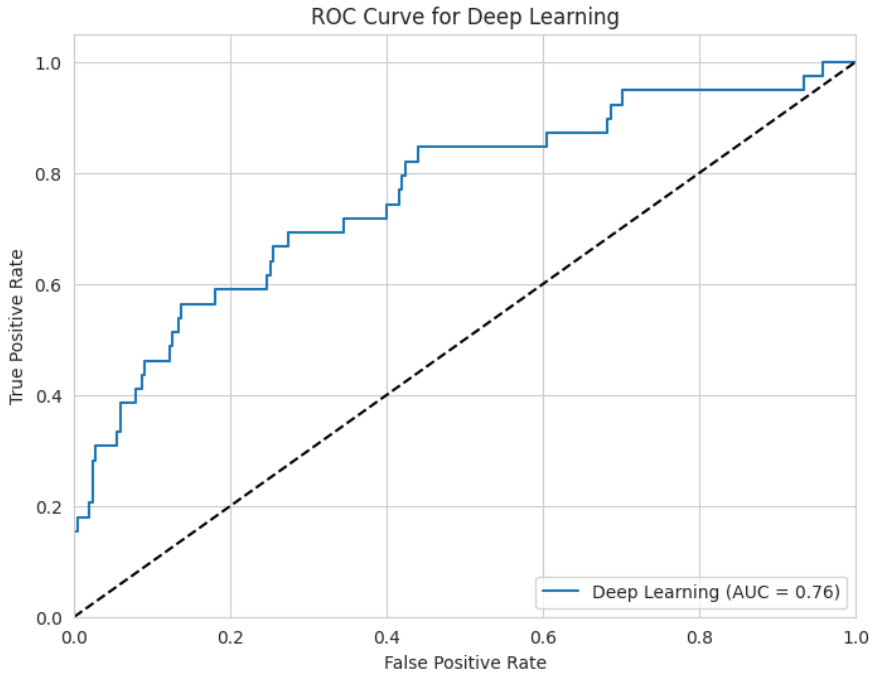


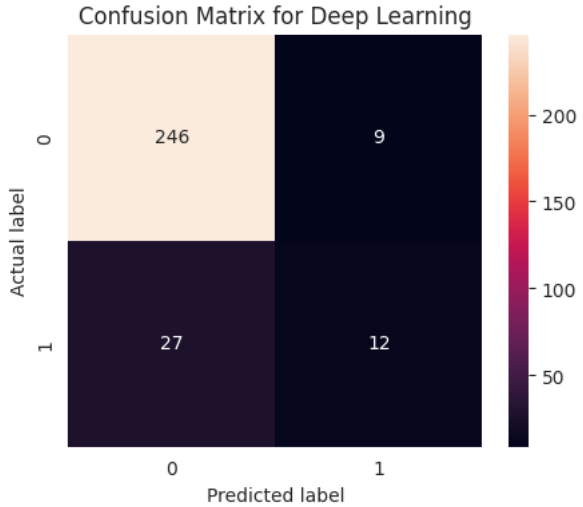
3) Neural Networks

The model includes ‘adam’ as optimizer, ‘relu’ and ‘sigmoid’ as activation functions.

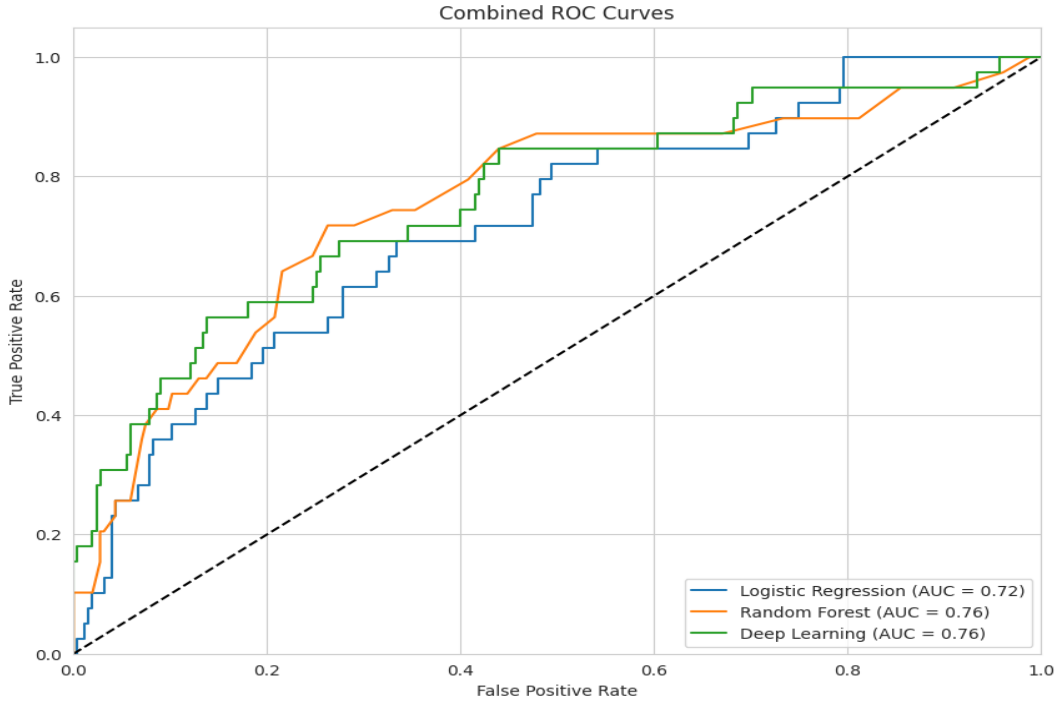
Data Partition: Training (70%), Validation(20%) and Test (10%)

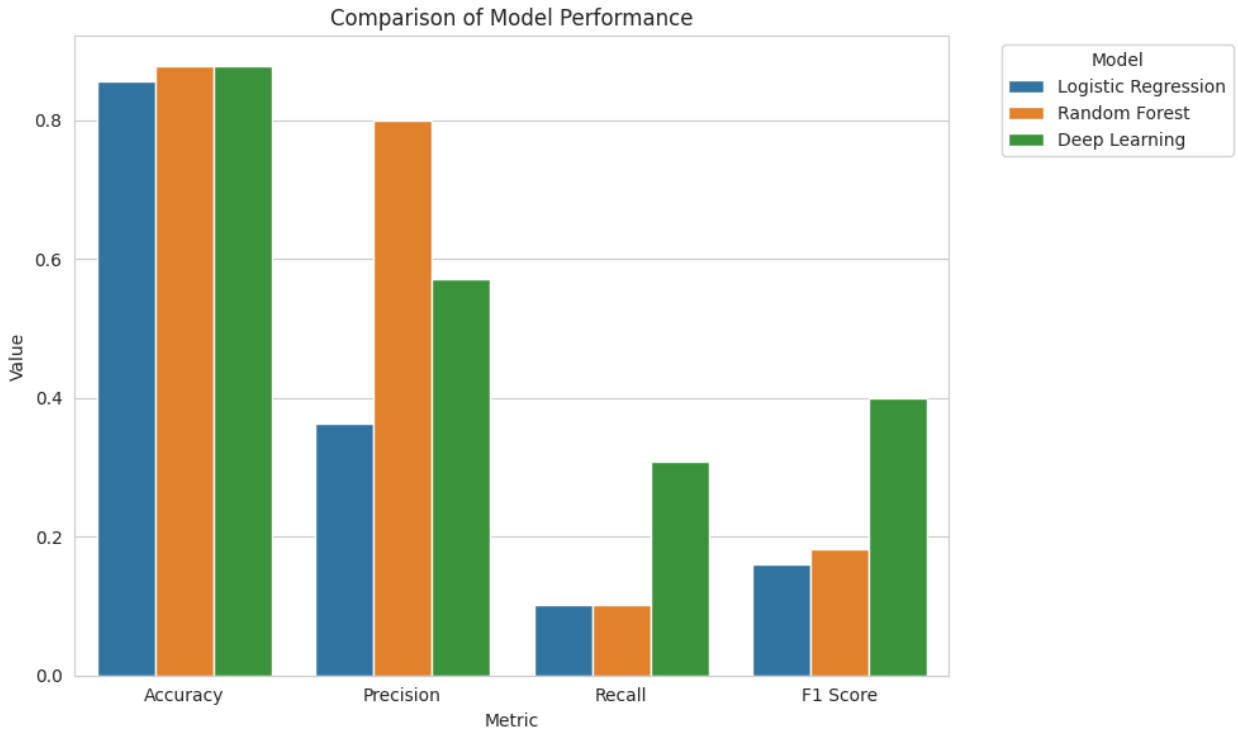
Accuracy: 87.7%, very similar to Random Classifier Model. This is probably due to the size of the dataset i.e., 1470 rows. Another reason is that the dataset is highly imbalanced as well.





Performance Comparison





Model	Accuracy	Precision	Recall	F1 score
Logistic Regression	0.85	0.36	0.10	0.16
Random Forest	0.87	0.80	0.10	0.18
Deep Learning	0.87	0.57	0.30	0.4

Based on the above results and ROC curves, Random Forest is the best model due to the high accuracy and F1 score.

Project Results

In this project, we aimed to predict employee churn using three machine learning models: Logistic Regression, Random Forest, and Deep Learning. Evaluation metrics—Accuracy, Precision, Recall, F1 Score and ROC curves were employed to assess each model's efficacy comprehensively. The comparison revealed that while the Random Forest and Deep Learning models both recorded high accuracy (0.87), the Random Forest model exhibited superior precision and an impressive F1 Score, signifying its balanced performance despite a shared lower recall with Logistic Regression.

The ROC curve analysis further highlighted the Random Forest model as the most effective for this prediction task, showcasing a favorable balance in detecting true positives while minimizing false positives. This equilibrium is essential for accurately identifying at-risk employees without triggering unnecessary interventions.

In conclusion, the Random Forest model emerged as the optimal choice for predicting employee churn, attributed to its balanced metric performance and high ROC curve AUC. This project demonstrates the significance of a multifaceted evaluation approach in model selection, providing a strong foundation for deploying machine learning solutions to enhance employee retention strategies effectively.

Impact of Project Outcomes

The impact of the project outcomes on organizational strategies for managing employee churn is significant. By accurately predicting which employees are more likely to leave, HR departments can tailor interventions more precisely, addressing the specific concerns and needs that drive turnover. This predictive capability allows for a proactive rather than reactive approach to retention, enabling organizations to enhance job satisfaction, align career progression opportunities more closely with individual aspirations, and foster a more engaging and supportive work environment.

Moreover, the insights gained from the project highlight the importance of a nuanced understanding of the factors contributing to employee churn. This understanding enables organizations to implement targeted strategies that not only reduce turnover but also contribute to a positive organizational culture and improved employee morale. As a result, businesses can expect to see a decrease in the costs associated with recruitment and training, an increase in productivity, and a more stable workforce that supports long-term organizational goals and success.

Ultimately, the project outcomes emphasize the transformative potential of leveraging machine learning in HR practices. By adopting models like Random Forest and Deep Learning for churn prediction, organizations can significantly improve their retention strategies, leading to a more resilient and high-performing workforce. This strategic advantage is crucial in today's competitive business environment, where retaining top talent is integral to innovation and growth.