Jyoti Kumari
Sarah Lee
Hrishi Salitri
Vannie Sung
Milan Vaghani

Employee Attrition Analysis

Description/Project Goals

Description:

We chose a Synthetic Employee Attrition Dataset from Kaggle. It is a robust resource designed to analyze and predict employee attrition within an organization. With 74,498 samples, this dataset provides a detailed view of various employee characteristics, Key variables include age, gender, job role, work-life balance, job satisfaction, performance rating, and factors such as monthly income, distance from home, and opportunities for career advancement. The dataset also encompasses employee perceptions of company reputation and recognition. By leveraging this data, we can uncover patterns and trends related to why employees leave or stay, assess the impact of various factors on attrition, and develop predictive models to identify employees more/less likely to stay. Ultimately, this analysis will help inform strategic HR policies and retention strategies, making it an invaluable tool for enhancing employee retention and improving organizational effectiveness.

Importance of Problem:

Attrition poses a major challenge for companies as it incurs substantial expenses for hiring and training new staff, hampers productivity, and can lower team morale and customer satisfaction. Frequent turnover also leads to the loss of crucial institutional knowledge and disrupts company culture, making it difficult to sustain a stable and skilled workforce. Effectively managing attrition is key to minimizing operational costs, boosting employee engagement, and achieving long-term organizational success. In our dataset we saw that the baseline, or average, attrition rate was 47%. In other words, 47% of the people in the dataset left their company.

Several factors influence employee attrition rates, including demographics, job satisfaction, work-life balance, performance ratings, compensation, career development opportunities, company culture, and remote work flexibility. For instance, employees dissatisfied with their jobs or lacking career growth opportunities are more likely to leave. Similarly, inadequate salaries, poor work-life balance, and a negative company culture can drive employees to seek employment elsewhere. Understanding which of these factors are the most essential for developing effective strategies to mitigate attrition.

By leveraging insights from attrition data, organizations can enhance their hiring and retention strategies. Targeted recruitment can identify candidates likely to stay with the company long-term, considering factors such as job fit, career aspirations, and cultural alignment.

Exploratory Analysis

During our EDA we explored all the variables included in the dataset to try to understand what looked to influence attrition the most. Out of the variables we saw the influence coming from: age, years at the company, job level, and work environment flexibility.

- Firstly, we observed that attrition rates decline as employees **age**, suggesting that older employees are more likely to remain with the company. The employee's age spanned from 18-59 and the average was about 39 years old. We saw that after 30, an employee is less prone to leave. [Fig. 1]
- Secondly, employees who have spent more **years at the company** tend to stay longer, indicating that the number of years at a company is a significant factor in retention. The average employment length was about 15.75 years with a standard deviation of 11.25 years. The minimum tenure was 1 year while the maximum reaches 51 years. This variation and range show a diverse workforce with both short- and long-term employees. [Fig. 2]
- From observing **job level,** we saw that the higher up an employee was at the company, the less likely you are to leave. Of those in a senior position, ~80% stayed, whereas those with entry positions had only 36% who stayed. [Fig. 3]
- The **work environment**, particularly the flexibility of remote work options, also looked like a crucial factor. Companies offering more flexible work-from-home arrangements have lower attrition rates of around 24%, underscoring the impact of a supportive work environment on employee retention. [Fig. 4]
- **The number of promotions** did not seem to have a big impact on attrition, until it got to a certain threshold. 75% of Employees stayed that had at least 3 promotions. This is also correlated with years at the company since you must stay at the company for a certain amount of time to gain promotions. [Fig. 5]
- The **Education level** of employee showed more education led to more loyalty. Employees with a bachelor's degree exhibit the highest attrition rate with 49% leaving the company, while those with a PhD showed the lowest attrition rate of 24%. [Fig. 6]
- In **marital status** We saw that the single population tends to leave company frequently whereas the married population tends to look for more stability hence they stayed in the company. [Fig. 7]
- Similarly, the greater **number of dependents** a person has, the less likely they are to leave the company. The average number of dependents per employee is approximately 1.65, with a standard deviation of 1.56, indicating a wide range of family sizes. Note, 25% of employees have no dependents. [Fig. 8]
- We saw a positive correlation between **company reputation** where we saw the higher the rating for the company, more employees stayed. [Fig. 9]
- **Gender** had effect too, with female employees have a higher attrition rate compared to male employees. [Fig. 10]
- Lastly, employees who perceive their **work life balance** more positively are more likely to remain with the company. [Fig. 11]

In summary, the analysis identified several key factors influencing employee attrition, including age, tenure, job level, work environment flexibility, education level, marital status, dependents, company reputation, gender, and work-life balance. By addressing these factors, companies can develop strategies to improve employee retention and create a more stable and satisfied workforce.

Solution and insights

The naive prediction for the data is 53% accurate. The overall mean of people that stayed in their respective companies was 53%, so naively if we predict everyone to stay in their company, we would be right about half the time. The goal of our modeling was to improve on this error and to more accurately predict if someone will leave their company or not.

K-Nearest Neighbors

In approaching the classification problem of predicting employee attrition, a K-Nearest Neighbors model was performed. The initial KNN model achieved an accuracy of 67.3% on the test data set. After incorporating cross validation to find the optimal number of neighbors (k) of 95, the accuracy improved to 71.4%. Further enhancement involved feature engineering. Polynomial interaction terms were created, and SelectKBest with ANOVA F-values was used to select the top 20 most relevant features. The final model achieved an accuracy of 71.8%. Feature importance revealed that both individual features and interaction terms were significant with job related and demographic factors in predicting attrition.

Logistic Regression

The formula for the logistic regression model used all variables, making sure they encoded the categorical predictors correctly through the design matrix. No special binning or feature engineering was used in this, so as not to complicate the approach. The model performed at a 65% accuracy rate on holdout data, with precision of 66% and recall of 70%.

Naive Bayes

To prepare for Naïve Bayes analysis, continuous variables were binned into quantiles or specific intervals: Age, Monthly Income, Distance from Home, and Company Tenure. These variables were divided into 5 quantiles with an equal number of observations. This normalization helps in comparing separate groups more effectively. The variable Years at Company was also binned by specific intervals to reflect key career stages and provide insights into tenure-related trends. These transformations make it easier to analyze how these variables impact other factors in the dataset.

The Naïve Bayes model trained on these binned variables and remaining categorical variables achieved an accuracy of 75%, with precision and recall values also at 75%.

Random Forest

The random Forest model was constructed by binning Age and Years in Company and creating a Job Level-Income feature. After optimizing the model with 1500 trees and 8 features, it achieved an RMSE of 75.2% and 76% Precision and Recall. Key features were Age, Years at Company, Job Level, and Marital Status.

The variable importance plot [fig. 13] highlighted the significance of Age, Years at Company, Job Level, and Marital Status in predicting employee outcomes. These features were crucial in determining employee behavior and contributed significantly to the model's overall performance.

Gradient Boosting

Variables mentioned above were also used initially in binning techniques, but predictive accuracy did not improve (decreasing very slightly, but almost negligent). Because of this the final boosting model did not have binned variables. After cross validation, 200 estimators, or trees, were used in the boosting iterations with a max depth of 4 for each tree. It was 75.7% accurate, so not much improvement was made from the simpler models, however we found important insights from this model to enhance our understanding.

From the variable importance chart [fig. 12] we can see the top 20 most important variables for predicting attrition in model. We are only looking at the top 20 because anything after this showed little to no importance in the model. Things like Marital Status, Job Level, Work Environment, and Work-Life Balance stand out here. From this we re-ran the model with only the 10 most important variables. Our accuracy decreased by 1.5% (74.2% accurate). To further explore important variables, we ran the model using only the four most important predictors (mentioned above). With only these four predictors, our accuracy was at 73%. Although this was a reduction from before, it further shows that the Marital Status, Job Level, Work Environment, and Work-Life balance of an employee has the biggest impact on their attrition. The boosting model gained nearly all its predictive power from these predictors.

Appendix: Supplementary Images

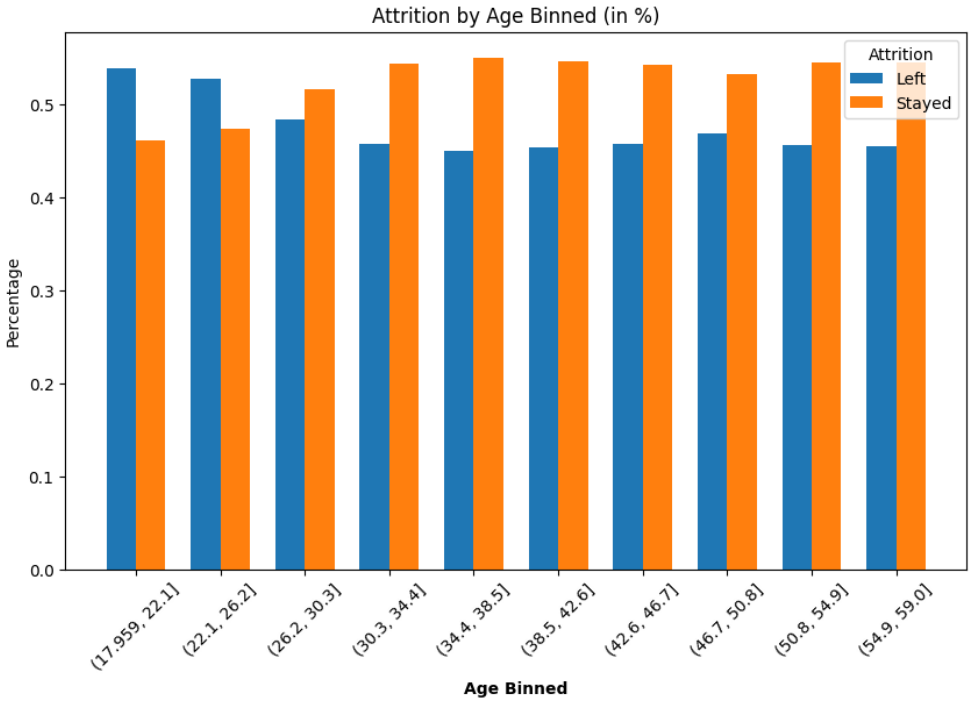


Fig.1: Bar chart Age binned (in %)

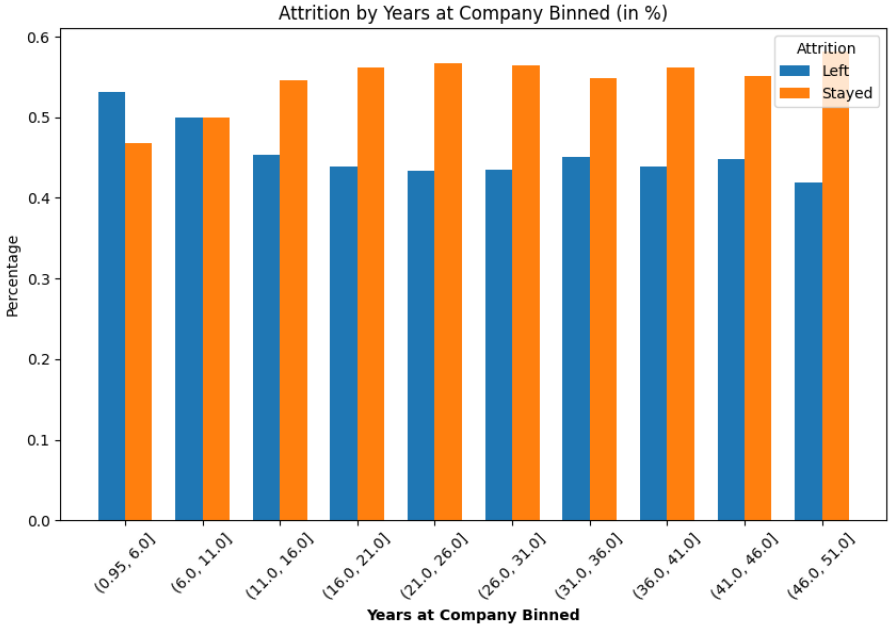


Fig.2: Bar chart Years at company (in %)

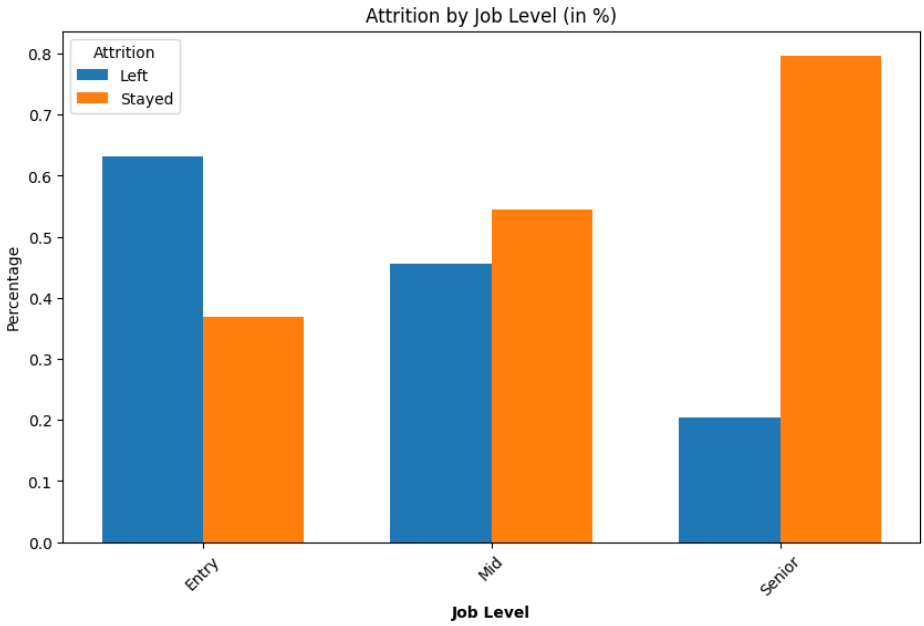


Fig.3: Bar chart Job level (in %)

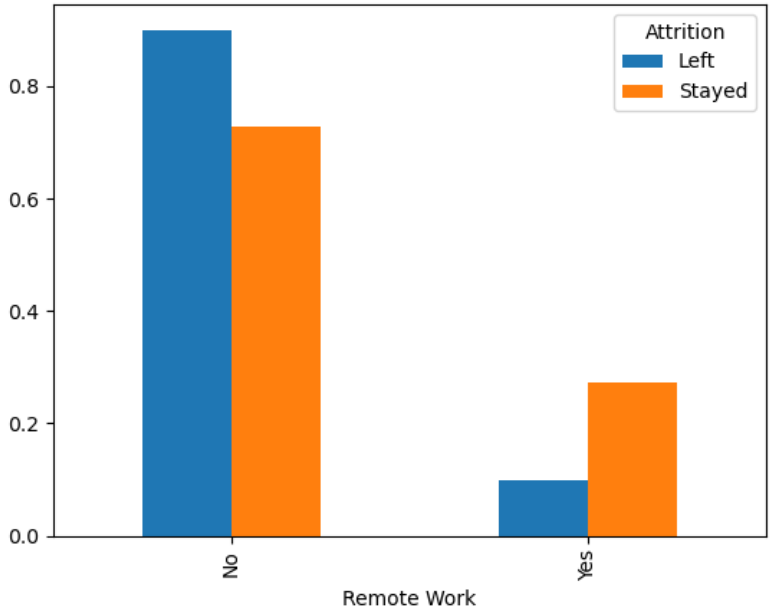


Fig.4: Bar chart Remote work (in %)

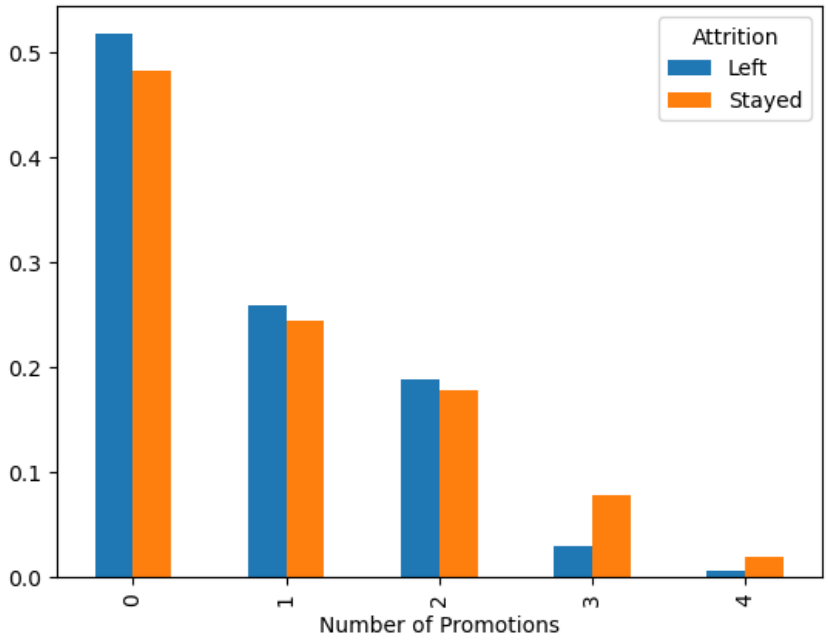


Fig.5: Bar chart Number of promotions (in %)

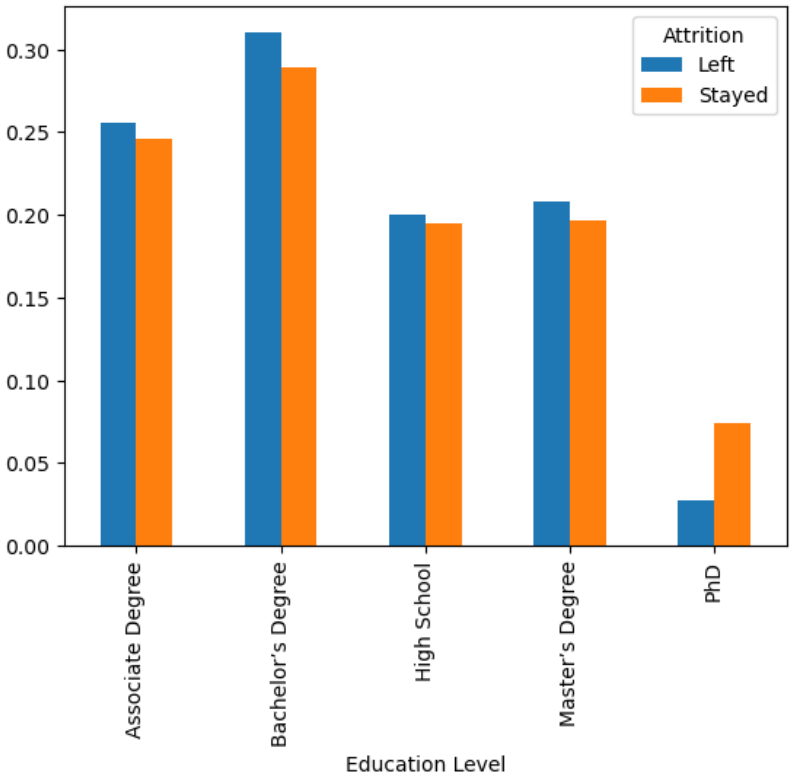


Fig.6: Bar chart Education level (in %)

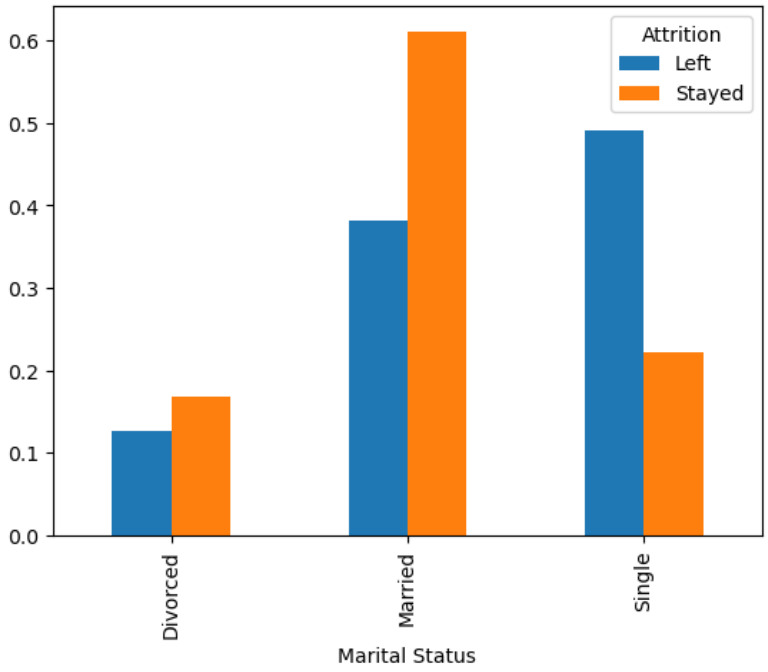


Fig.7: Bar chart Marital status (in %)

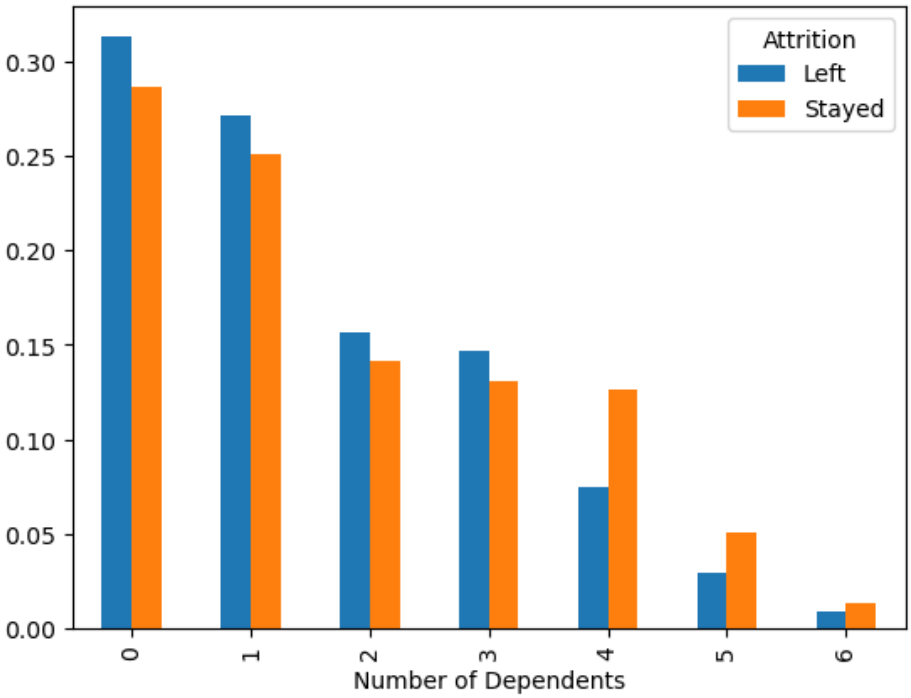


Fig.8: Bar chart Number of Dependents (in %)

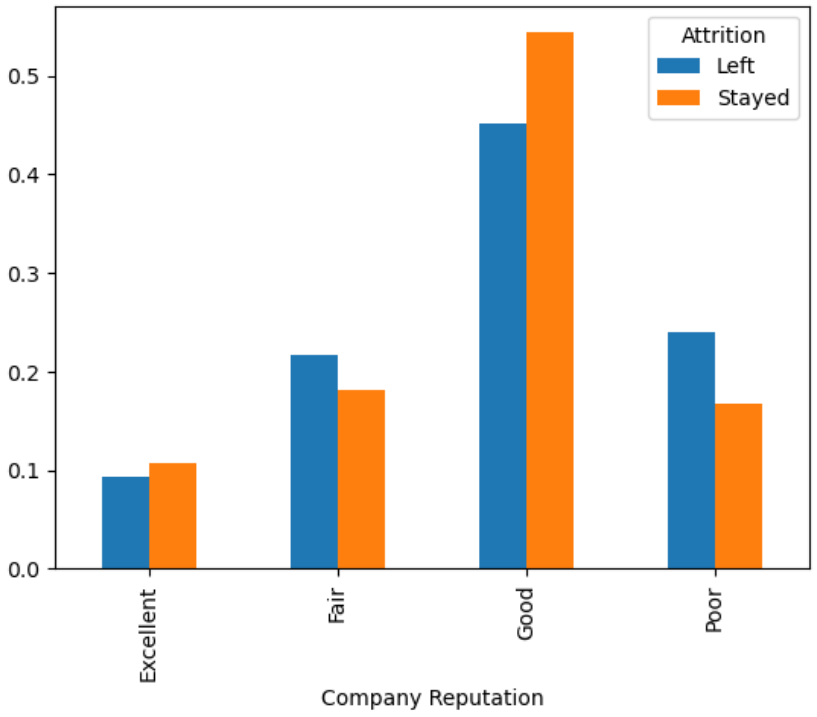


Fig.9: Bar chart Company Reputation (in %)

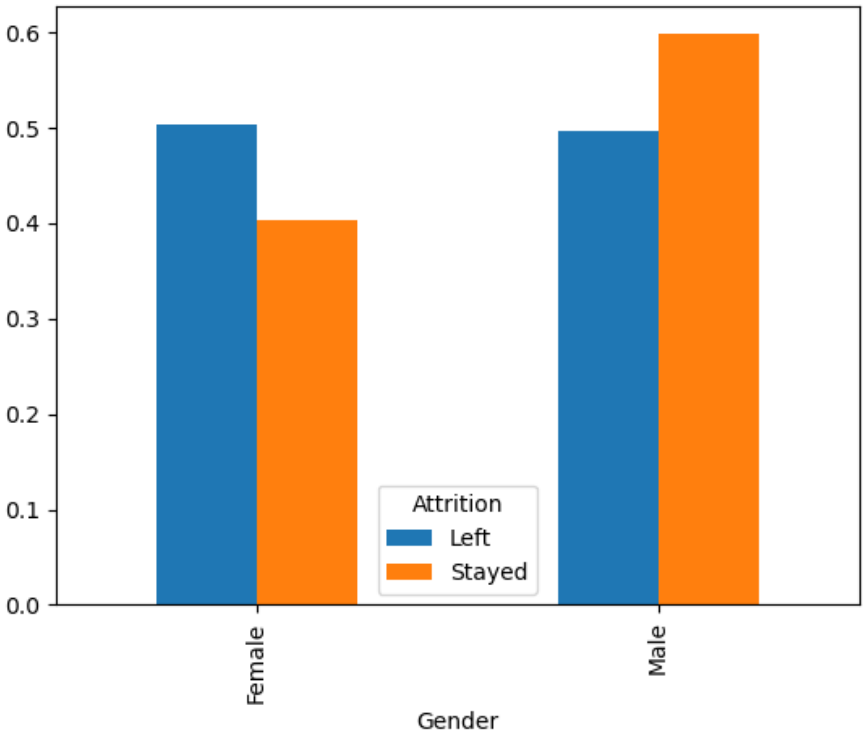


Fig.10: Bar chart Gender (in %)

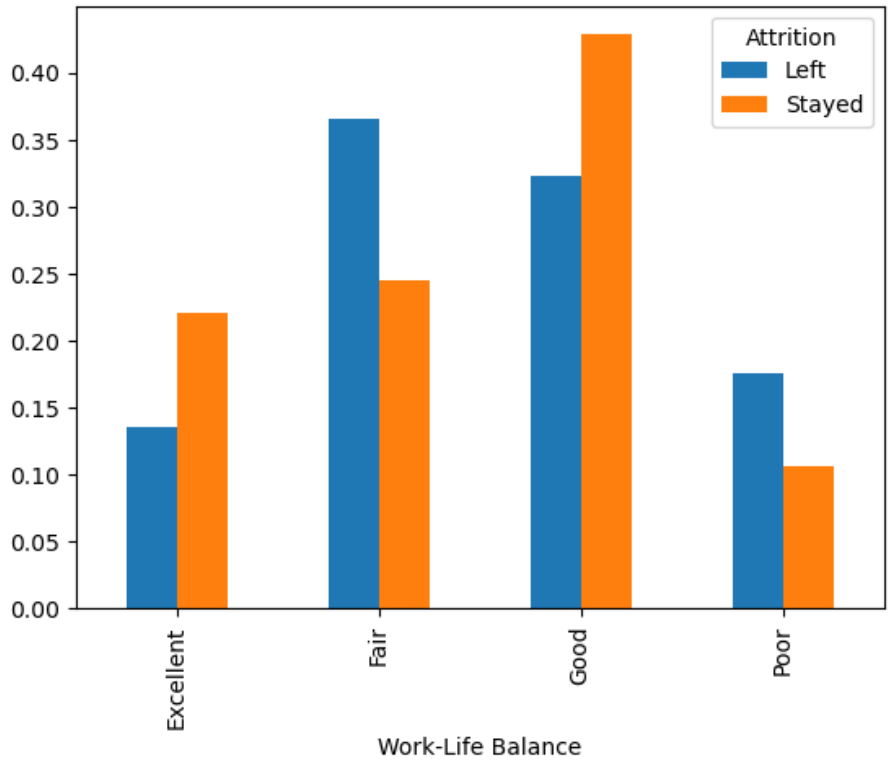


Fig.11: Bar chart Work-Life Balance (in %)

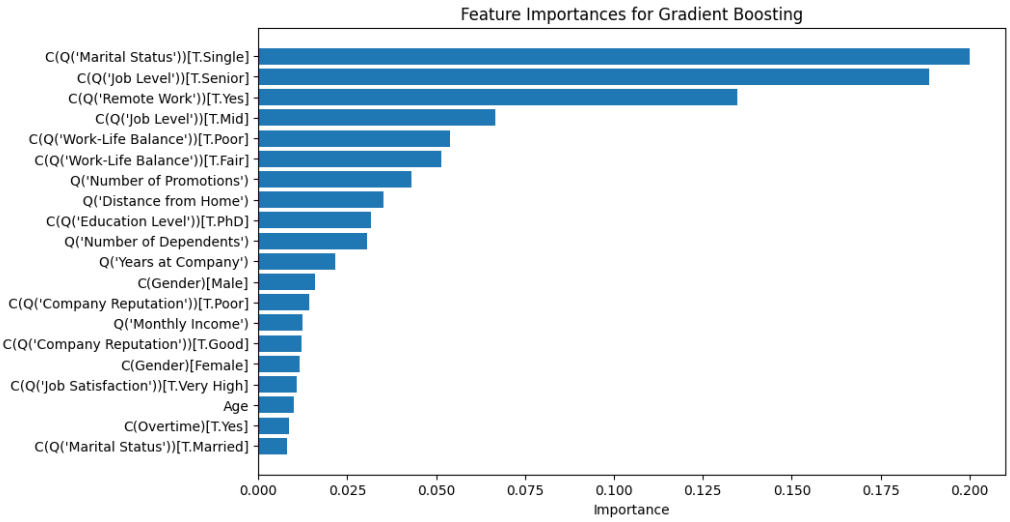
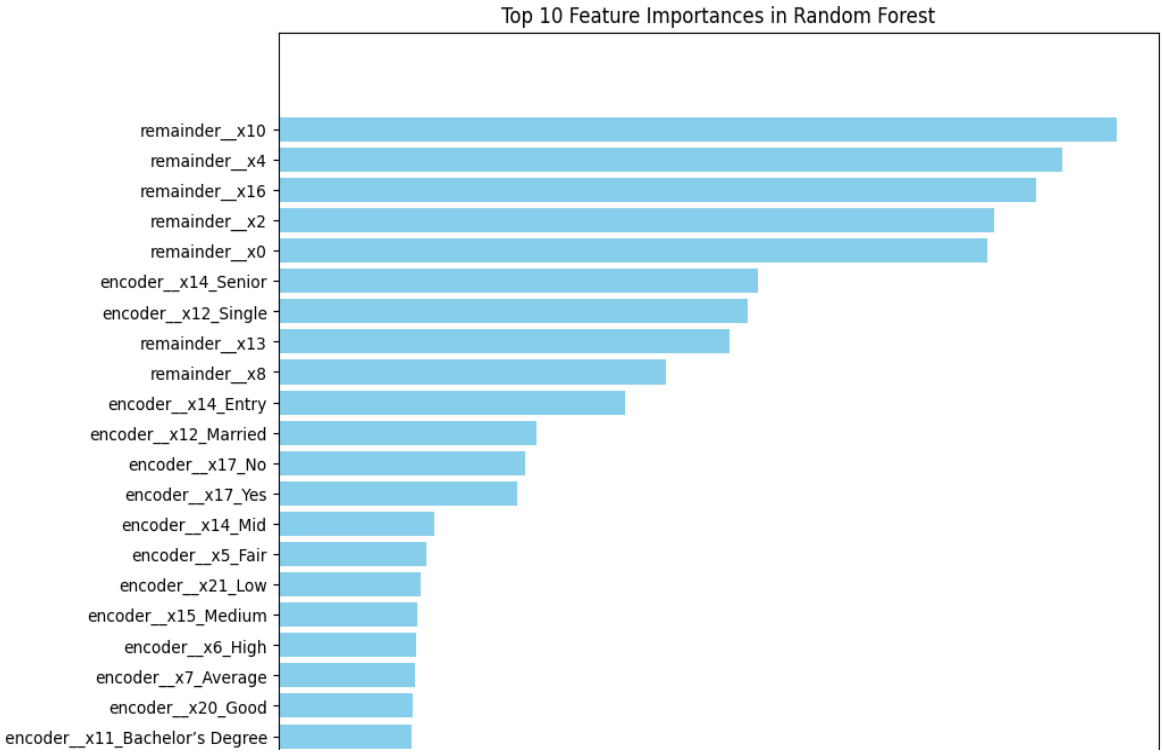


Fig.12 Feature importance Gradient Boosting Model



Fig,13 Feature importance Random Forest Model

References

1. MIS S381N- Data Science Programming Python codes
2. Kaggle Employee Attrition Classification Dataset
3. ChatGPT