FINDINGS ON ATTRITION

BRAIN AND COMPANY CONSULTANCY

EXECUTIVE SUMMARY

- After evaluating Logistic Regression, K-nearest neighbors, and Random Forest models, I focused on Logistic Regression and Random Forest due to their highest accuracy scores. Random Forest achieved an impressive 99% accuracy, but this might lead to overfitting issues on new data. Logistic Regression, with an 85% accuracy, proved reliable and not overtrained. Despite having more false negative predictions, it aligns well with our focus on attrition rather than non-attrition. Moreover, its performance corresponds strongly with insights from the correlation matrix chart, enhancing our confidence in its predictive ability.
- Featured in training the models included: Age, business travel, department, gender, marital status, monthly income, num companies worked, percent salary hike, stock option level and total working years, years at company , job involvement /expansion
- I performed 20 percent test and 80 train split.
- Recommendation is to go with the Logistic Regression because reliable, not overtrained and easier to understand.

EXECUTIVE MODEL RECOMMENDATION

I recommend going with Logistic Regression because it's easy to understand and gives solid predictions. It might not be as perfect with scores, but it's reliable and won't let you down when faced with new data. Starting with logistic model would help us understand why people might leave and where we can improve. It's like having a trustworthy guide through the data jungle.

LOWER ATTRITION BY AGE AND MARITAL STATUS

- Age-Aligned Mentorship: Pair employees with mentors around the same age group to foster social connections beyond work.

- Optional Non-Work Events: such as trivia night, karaoke, game night, to build companionship within the company.

- Gym Memberships: Offer gym memberships to help lower attrition among employees in their 20s and 30s, as well as among single employees.

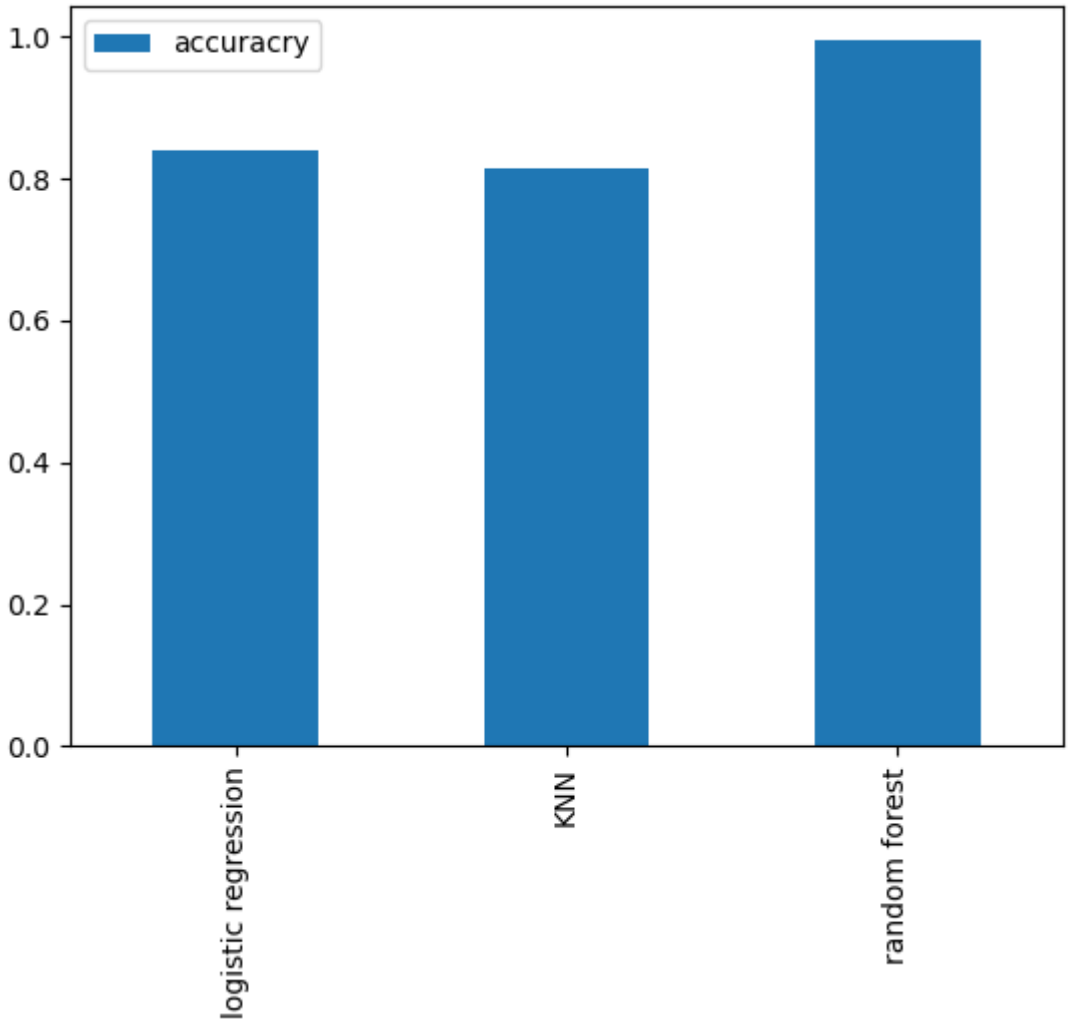
LOWER ATTRITION WITH EMPLOYES LESS THAN 2 YEARS

- Employees with less than 2 year at the company and employees with less than 2 years work experience tended to have higher attrition.
- More than 30 day training for those that score poor or average on performance review
- Enhanced mentoring and transitioning into role support
- Limit of 1 or 2 job “expansion” roles for employees with less than 2 years at the company.
- Job “expansion” should not be ranked, but optional opportunities that support employees can be beneficial. These can include engaging activities such as fundraisers or opportunities for upward mobility, which could lead to a higher role

RECOMMENDATION ON SURVEYS, EVALUATIONS, AND TRAVEL

- Let managers know they are encouraged to enter poor, average on employee performance rating.
- Let employees know their employee's name and Id is anonymous when feeling out forms to result in an honest rating.
- Let Business Travel percentage be known for position.
- Implement a survey to determine which qualified employees are interested in traveling and which are not. Assign frequent business travel opportunities to those who express a desire to travel.

WHAT MODEL IS MORE FIT?



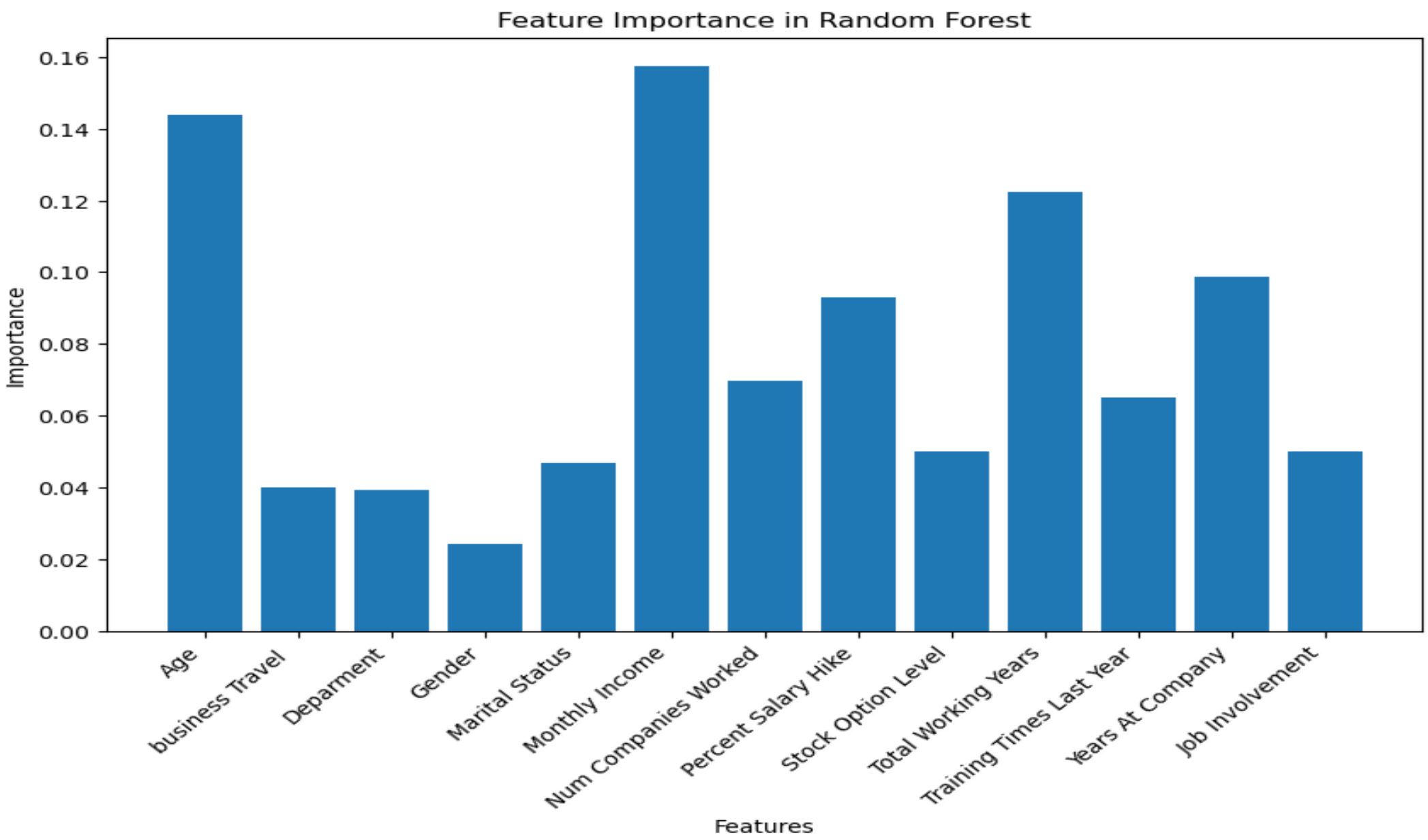
Accuracy: 0.99

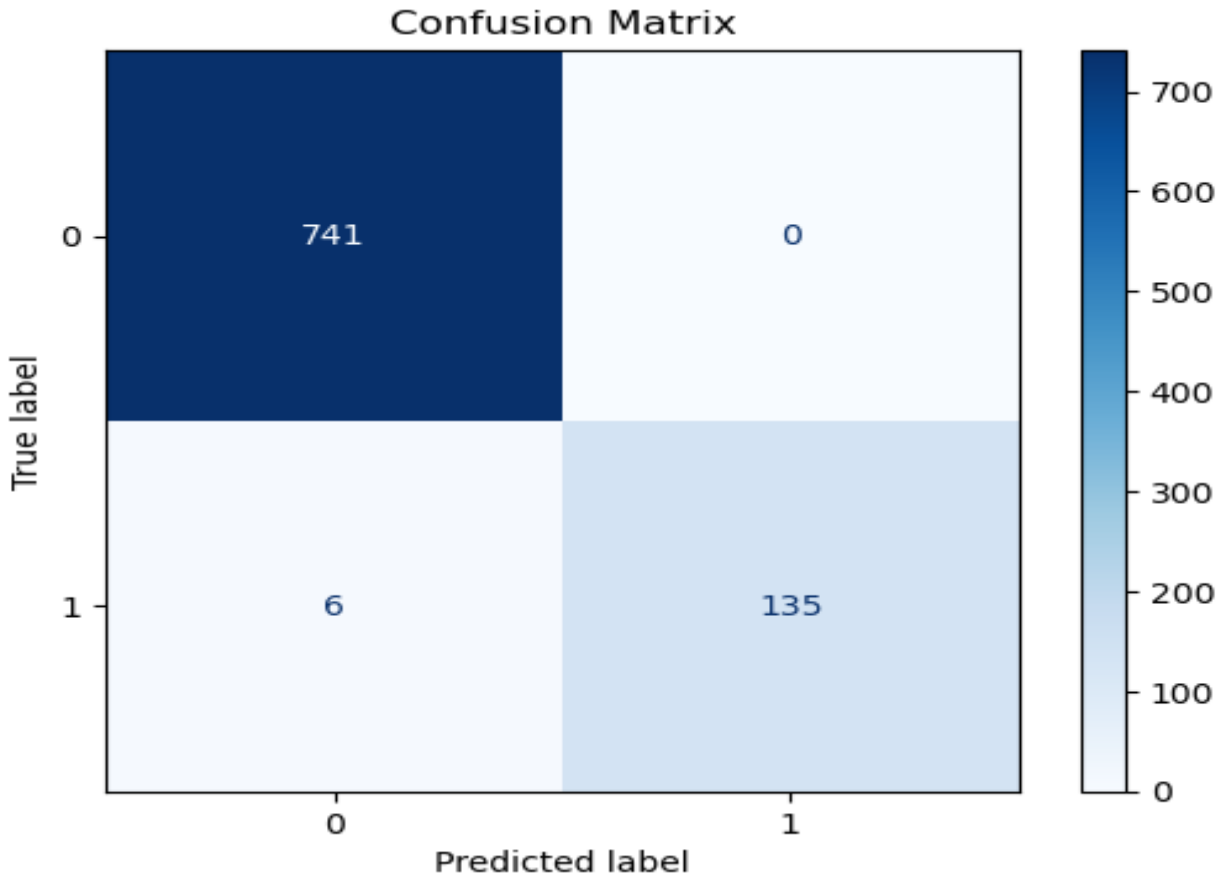
Classification Report:

	precision	recall	f1-score	support
0	0.99	1.00	1.00	741
1	1.00	0.96	0.98	141
accuracy			0.99	882
macro avg	1.00	0.98	0.99	882
weighted avg	0.99	0.99	0.99	882

RANDOM FOREST

CLASSIFICATION REPORT


Feature Importance in Random Forest
Importance
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Age
business Travel
Deparment
Gender
Marital Status
Monthly Income
Num Companies Worked
Percent Salary Hike
Stock Option Level
Total Working Years
Training Times Last Year
Years At Company
Job Involvement
Features



RANDOM FOREST

PREDICTED 6 NON-ATTRITION WRONG

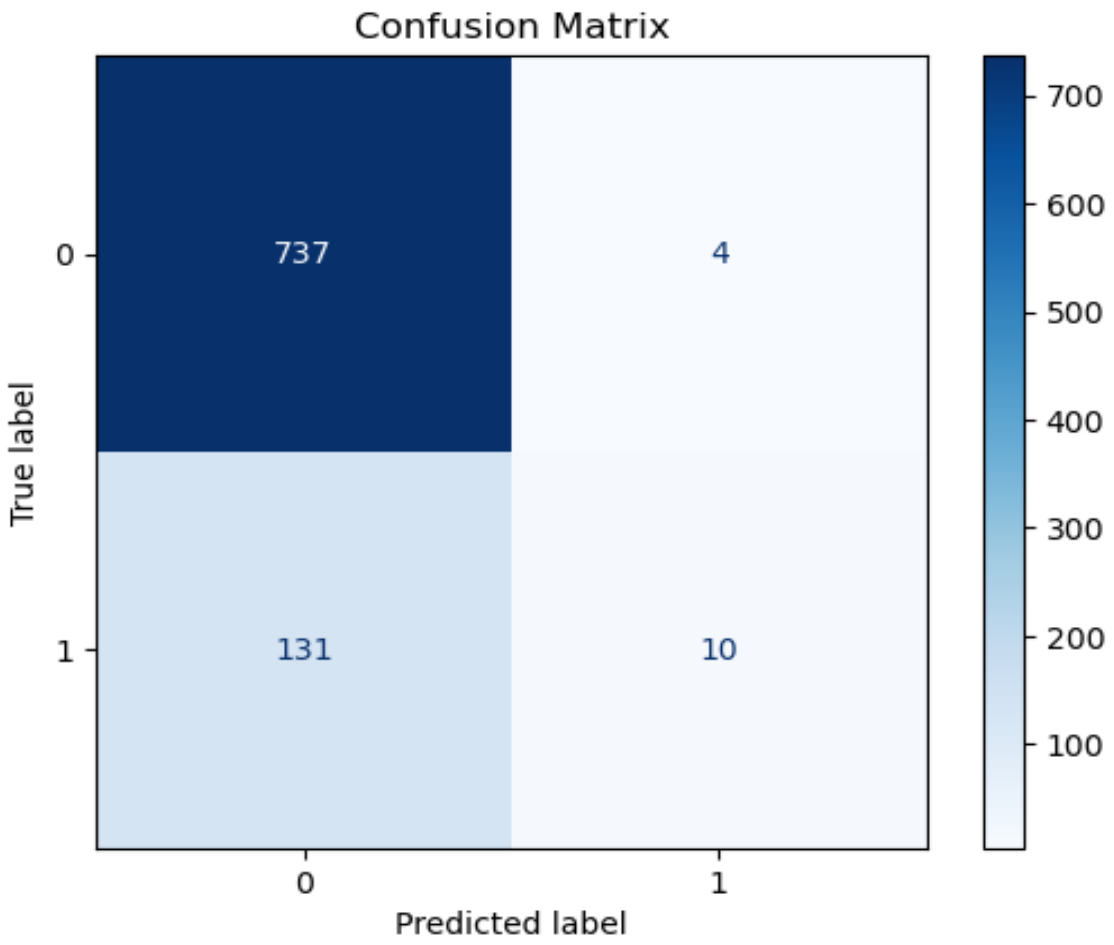
Accuracy: 0.8469

Classification Report:

	precision	recall	f1-score	support
0	0.85	0.99	0.92	741
1	0.71	0.07	0.13	141
accuracy			0.85	882
macro avg	0.78	0.53	0.52	882
weighted avg	0.83	0.85	0.79	882

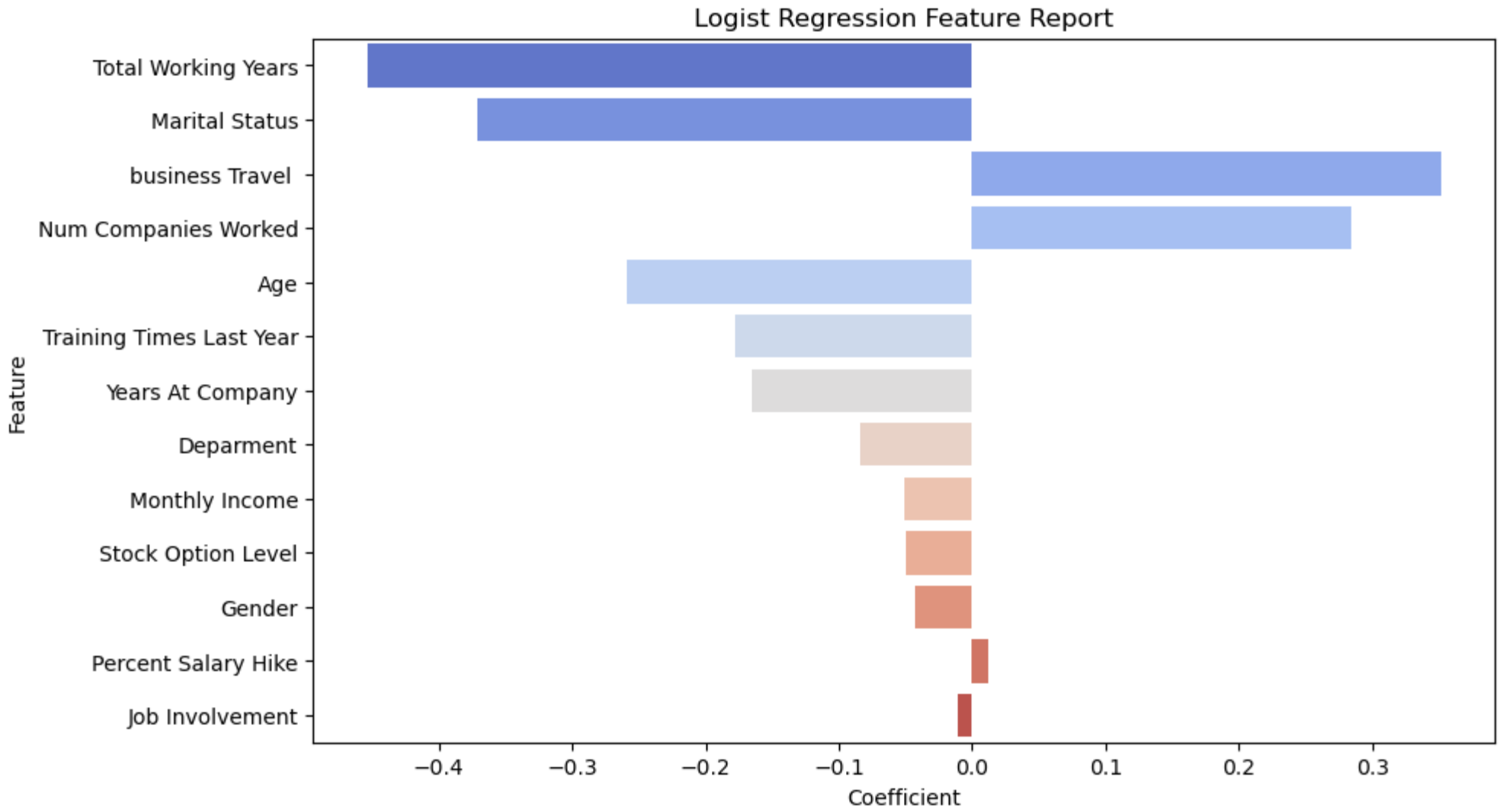
LOGISTIC REGRESSION

CLASSIFICATION REPORT



LOGISTIC REGRESS CONFUSION MATRIX

- PREDICTED: 4 EMPLOYE ATTRITION WRONG
- PREDICTED:131 EMPLOYE NON-ATTRITION WRONG


Logist Regression Feature Report
Total Working Years
Marital Status
business Travel
Num Companies Worked
Age
Training Times Last Year
Years At Company
Deparment
Monthly Income
Stock Option Level
Gender
Percent Salary Hike
Job Involvement
Feature
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
Coefficient

COMPANY STATS

- Total Employees Reported

  4,410

- Non Attrition Count/ Employees who stayed from previous year

  3,699

- Attrition count/ Employees who left from previous year

  711

- Attrition Rate

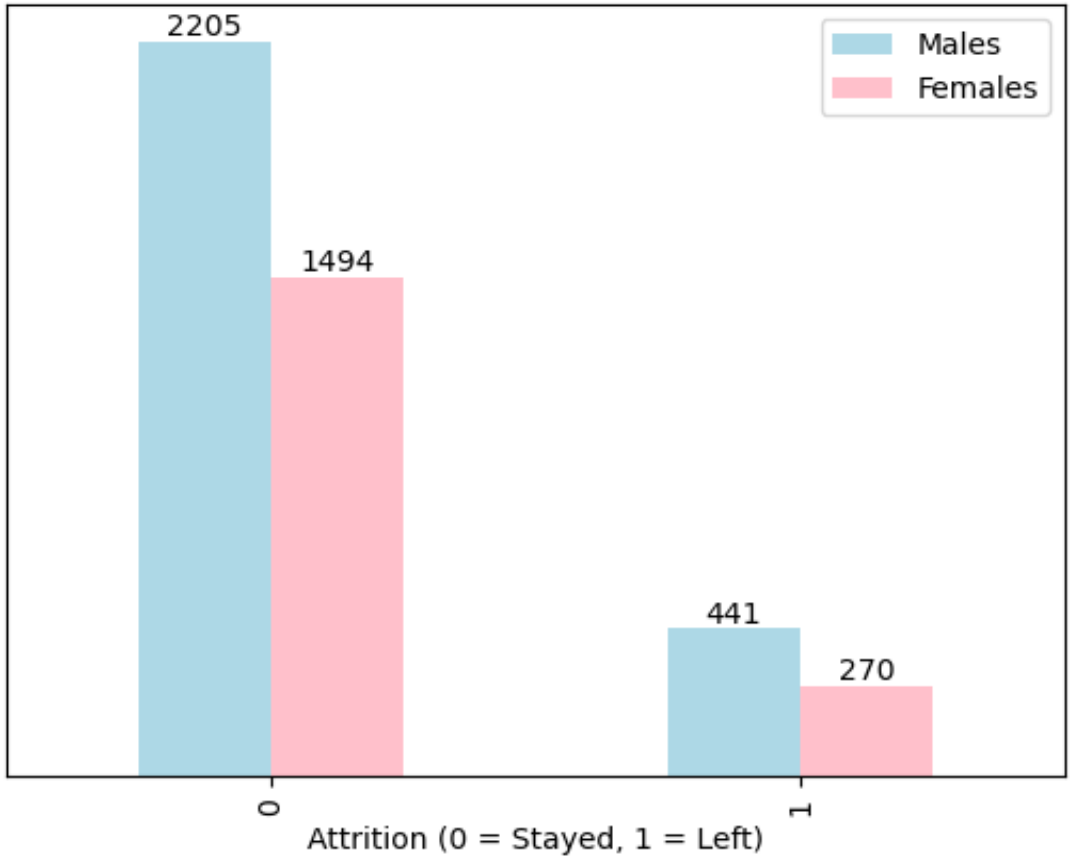
  16%

- Gender:

  Male = 2,656 or 60%

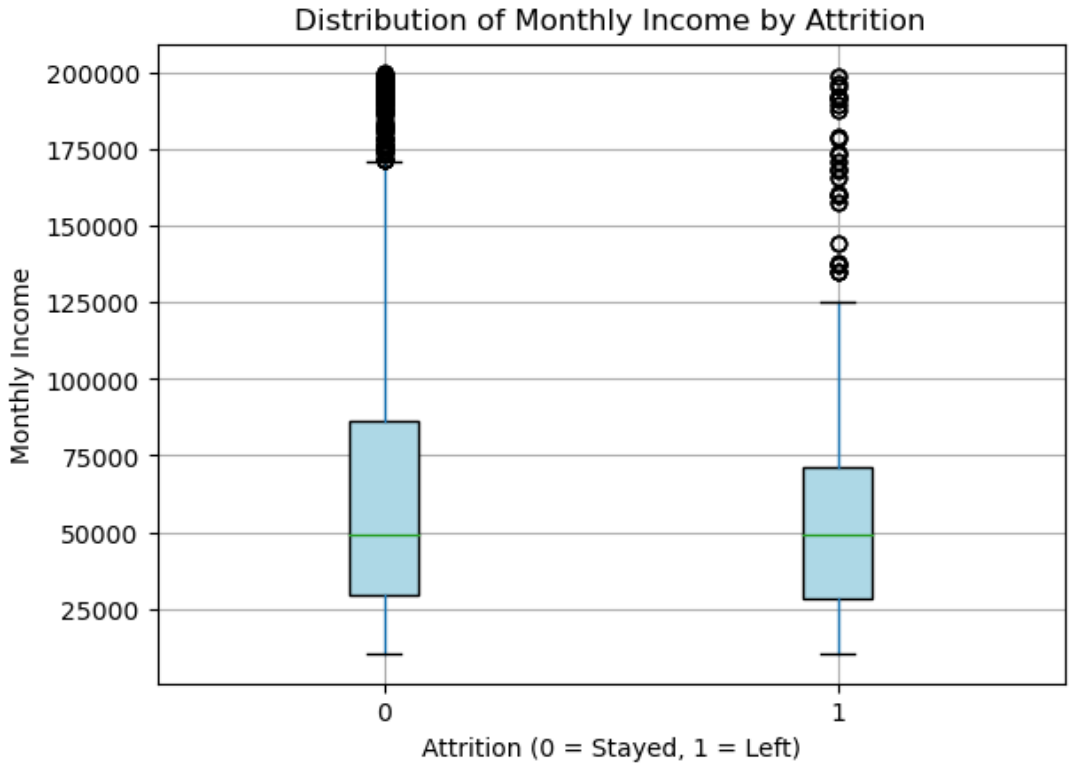
  Female = 1764 or 40%

ATTRITION AND NON- ATTRITION COUNT BY GENDER



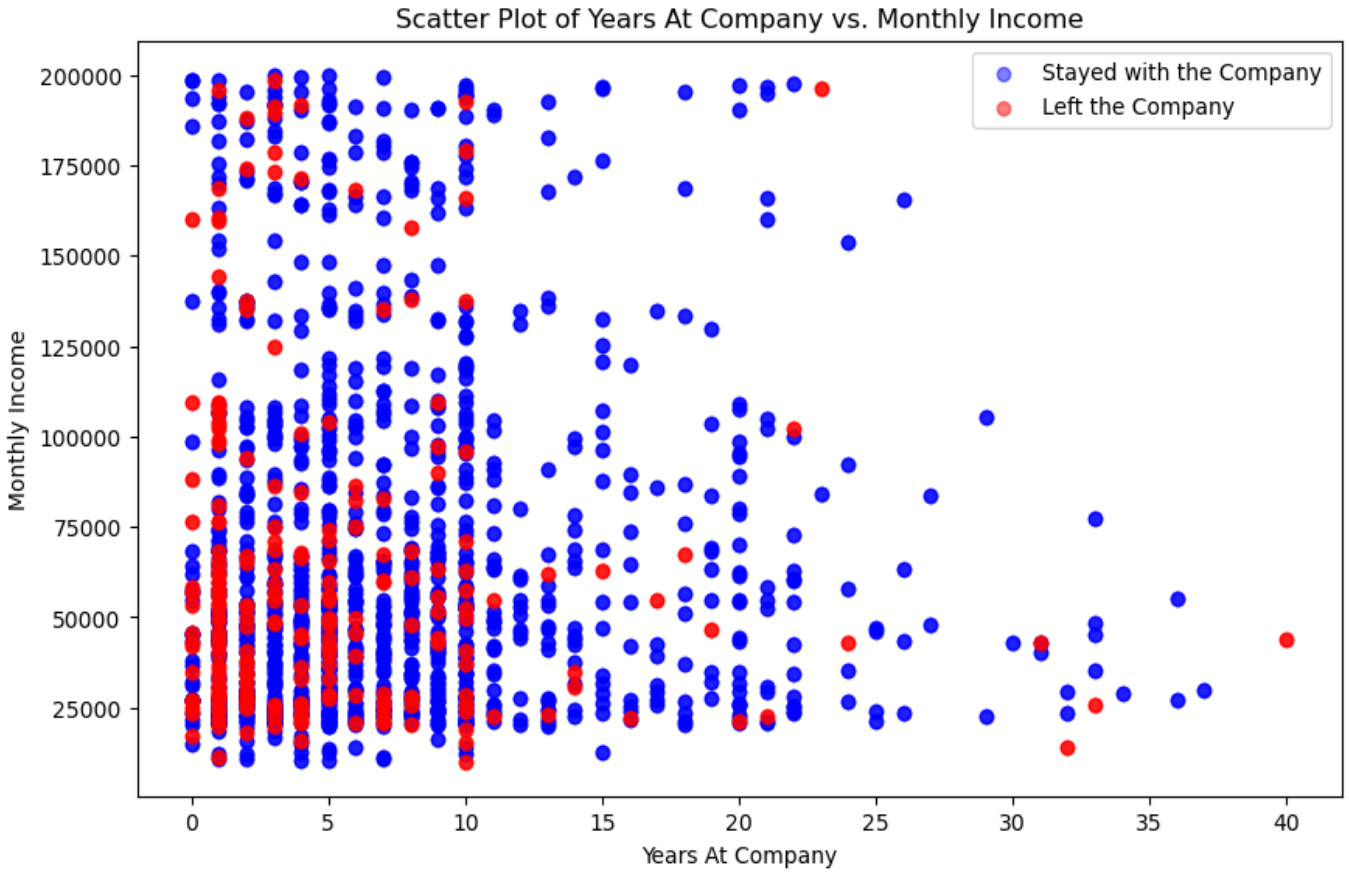
ATTRITION BY INCOME

- Using a box plot
- Mean for both is around 50,000 in Rupees
- Those that stayed have a larger monthly income range vs those who left.
- Outliers for no Attrition starts at 175,000
- Outliers for Attrition/left starts around 125,000



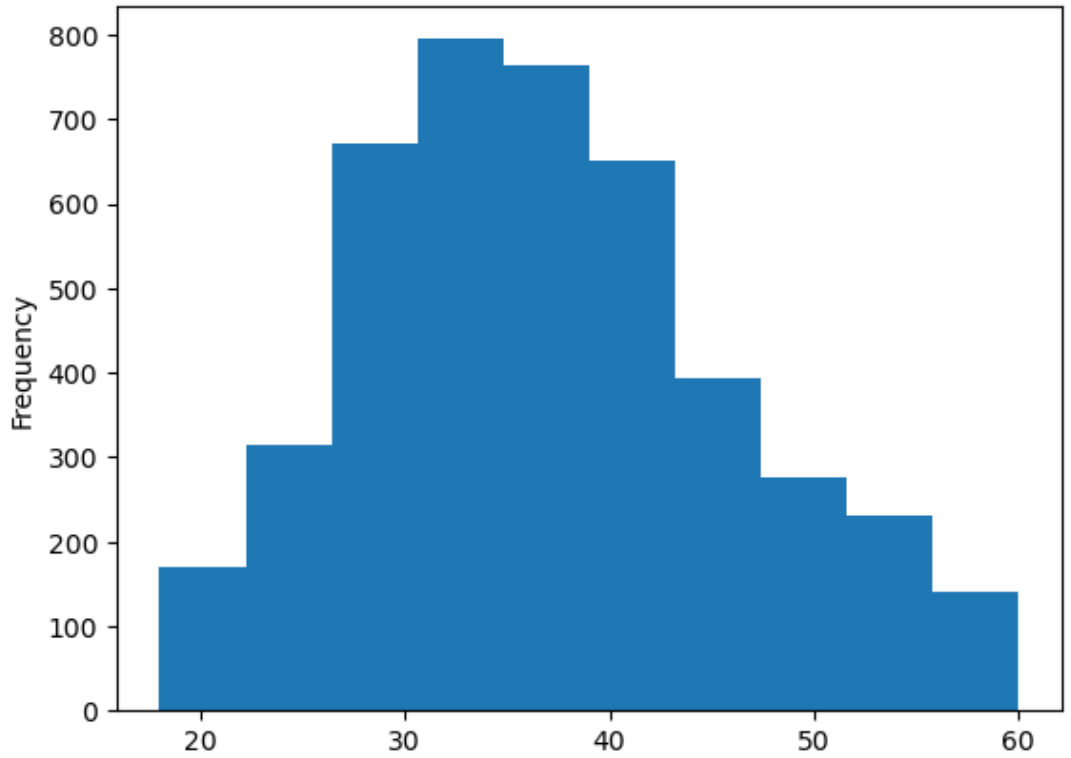
YEARS AT COMPANY AND INCOME BY ATTRITION VS NON-ATTRITION

- Using a scatter plot
- Scatter plot is scatter in a no real Lenier line
- Scattered randomly for both stated vs left
- Income scattered from around 25,000 to around 200,000.
- Most who left are within 25,000 to 110,000 monthly salary



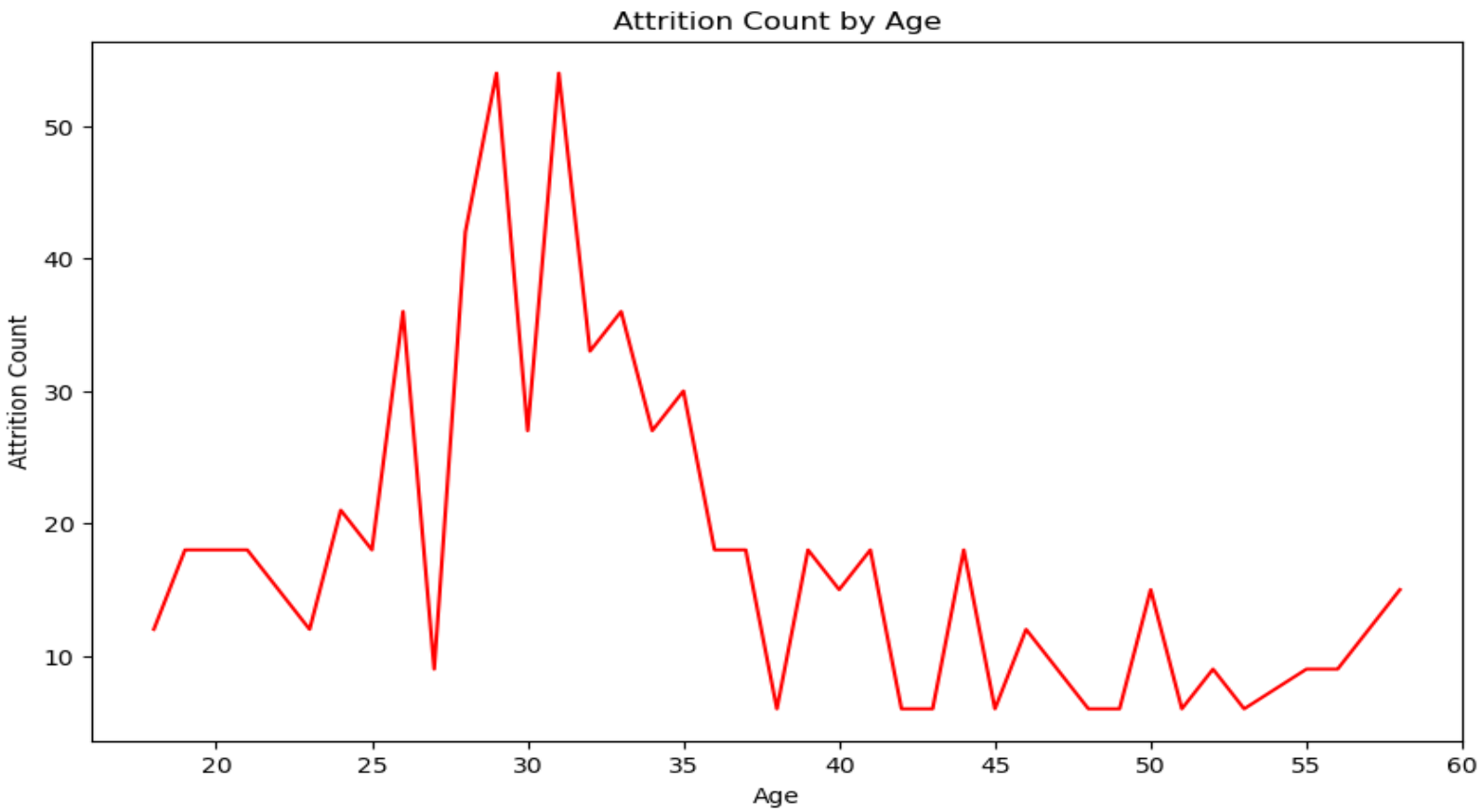
AGE RANGE AT THE COMPANY

- Using histogram
- Age: 20-60
- Highest Age range: 30s



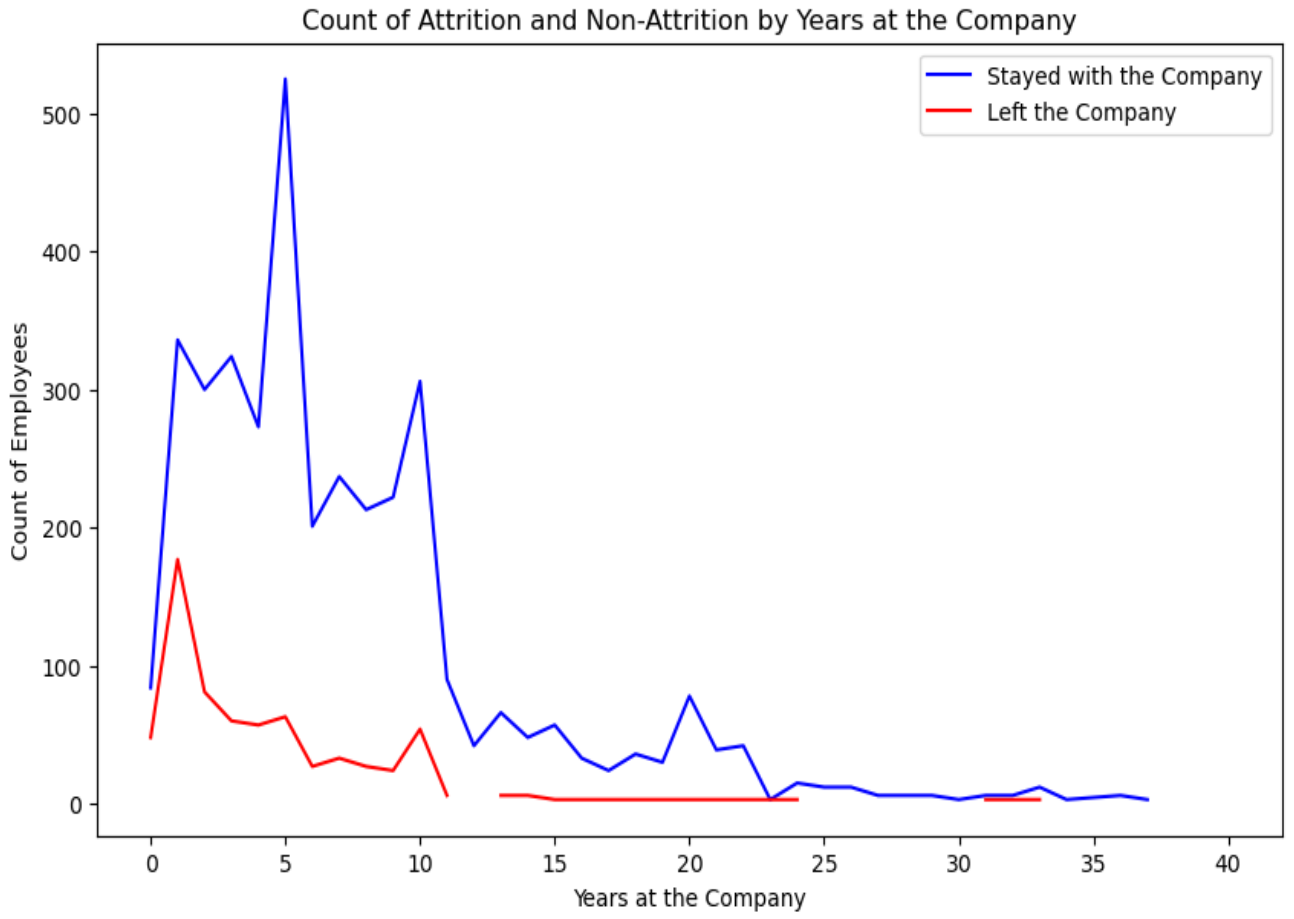
ATTRITION BY AGE

- Line Chart
- High Attrition between late 20s and mid 30s
- Attrition related to age drops after mid 30s



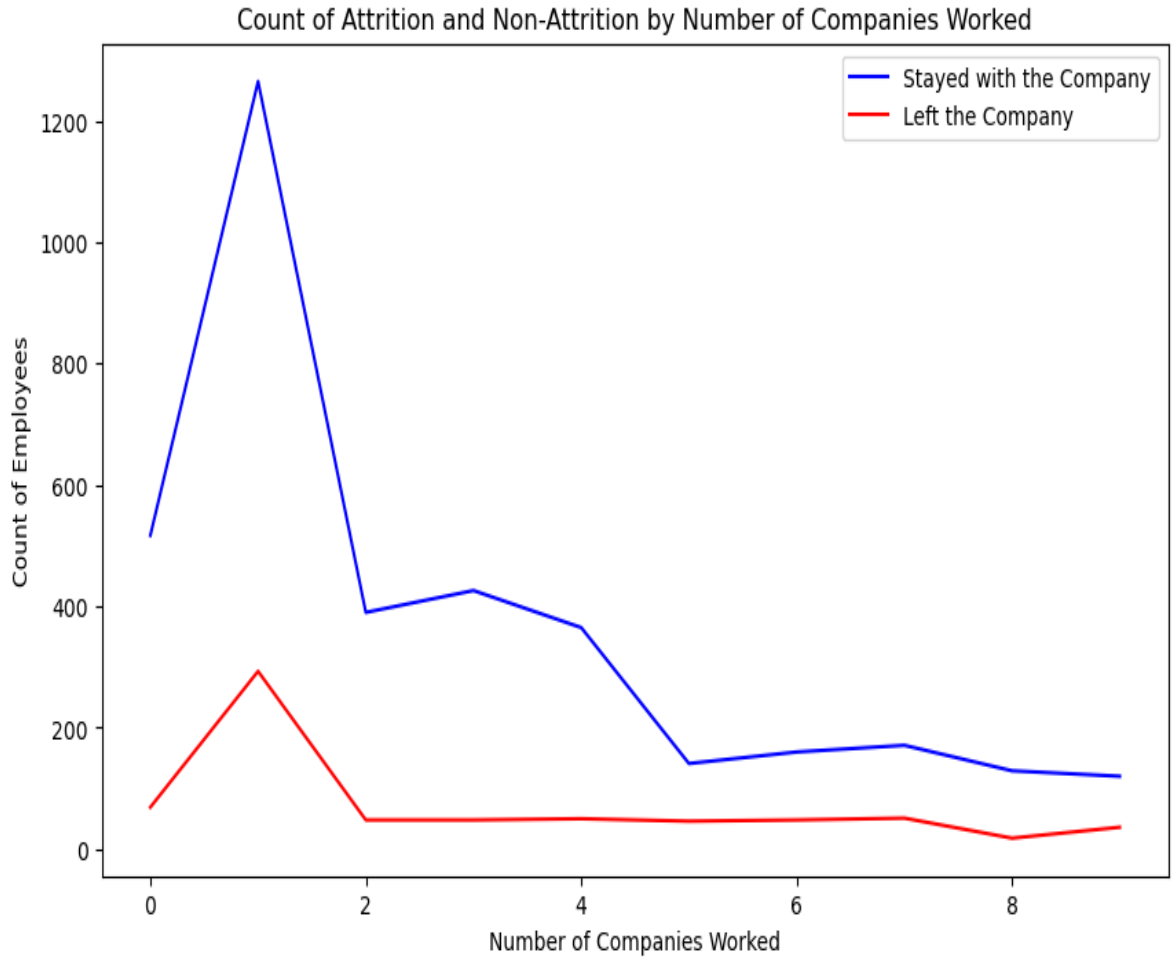
ATTRITION/NON-ATTRITION BY YEARS AT THE COMPANY

- Using Line Chart
- Most Employees who left are with the company for 1 year.
- After Employees first year attrition count lowers.



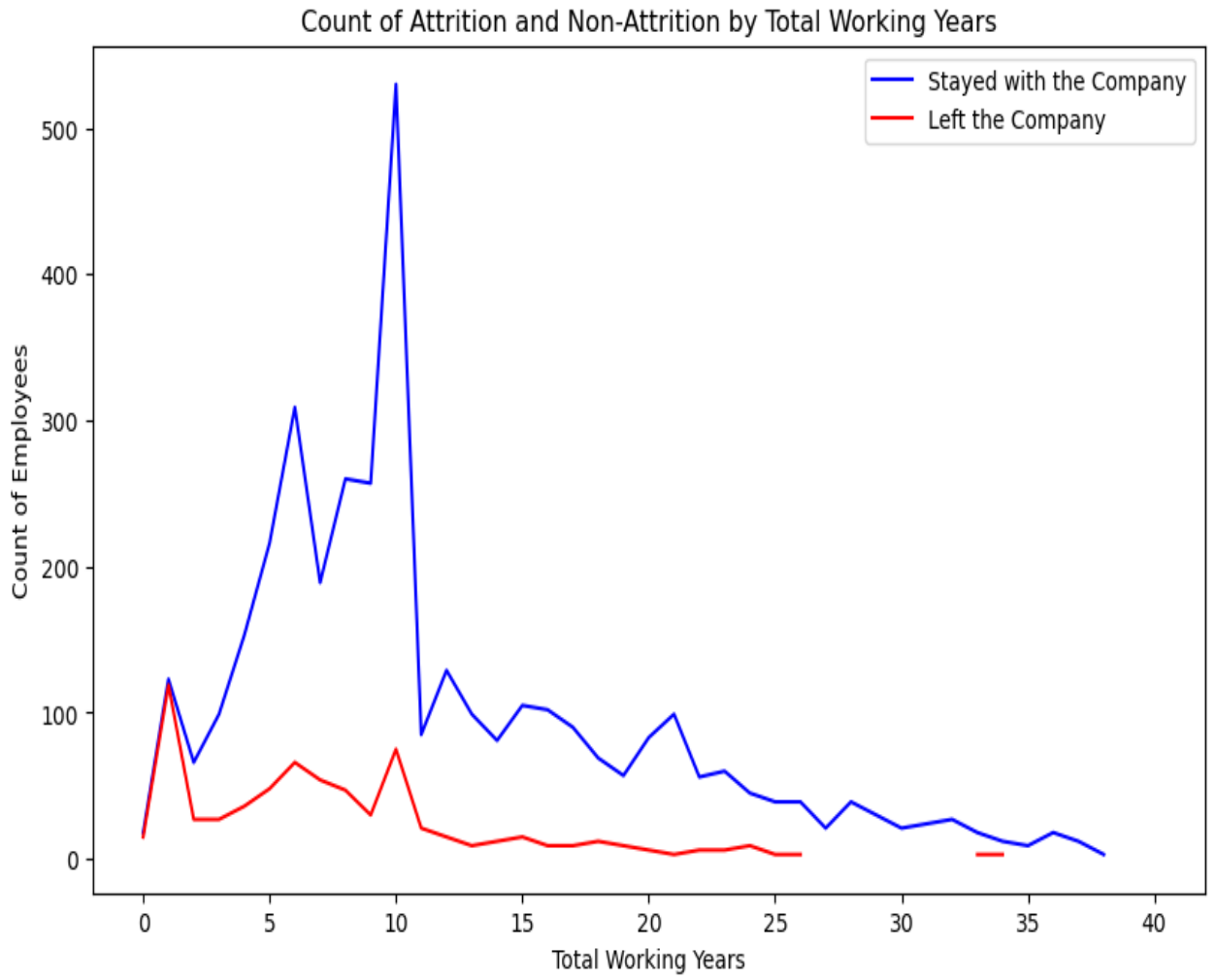
ATTRITION/NON-ATTRITION BY NUMBER OF COMPANIES WORKED

- Using Line Chart
- Employees who only worked at 1 company had the highest attrition rate compared to employees who work for more than 1 company.
- Employees who work for > 1 company resulted in lower attrition count



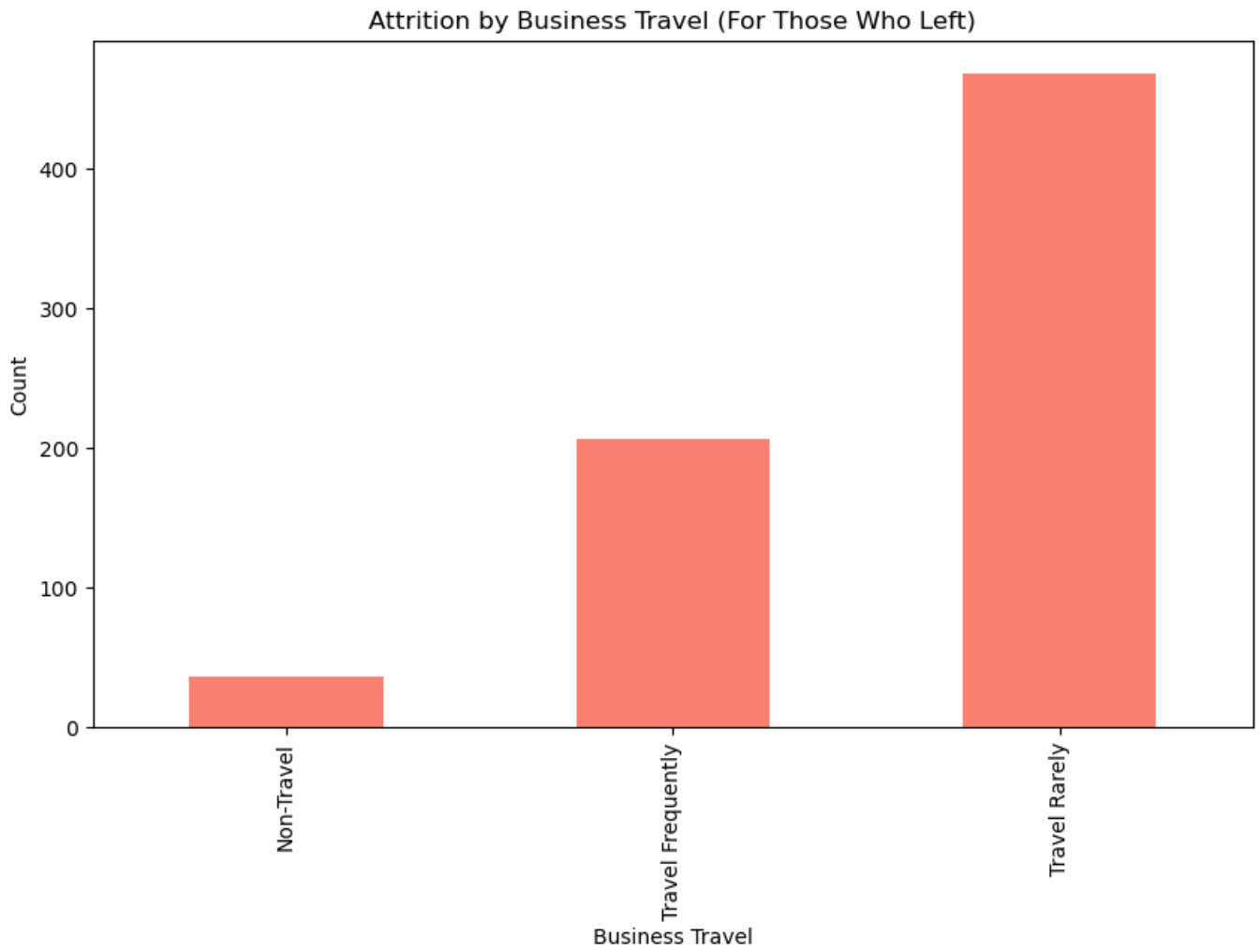
ATTRITION/NON-ATTRITION BY TOTAL WORKING YEARS

- Using Line Chart
- Employees whish around 1 year of total past work experience had the highest attrition count.
- Employees with more than 1 year of total work experience resulted in lower attrition count



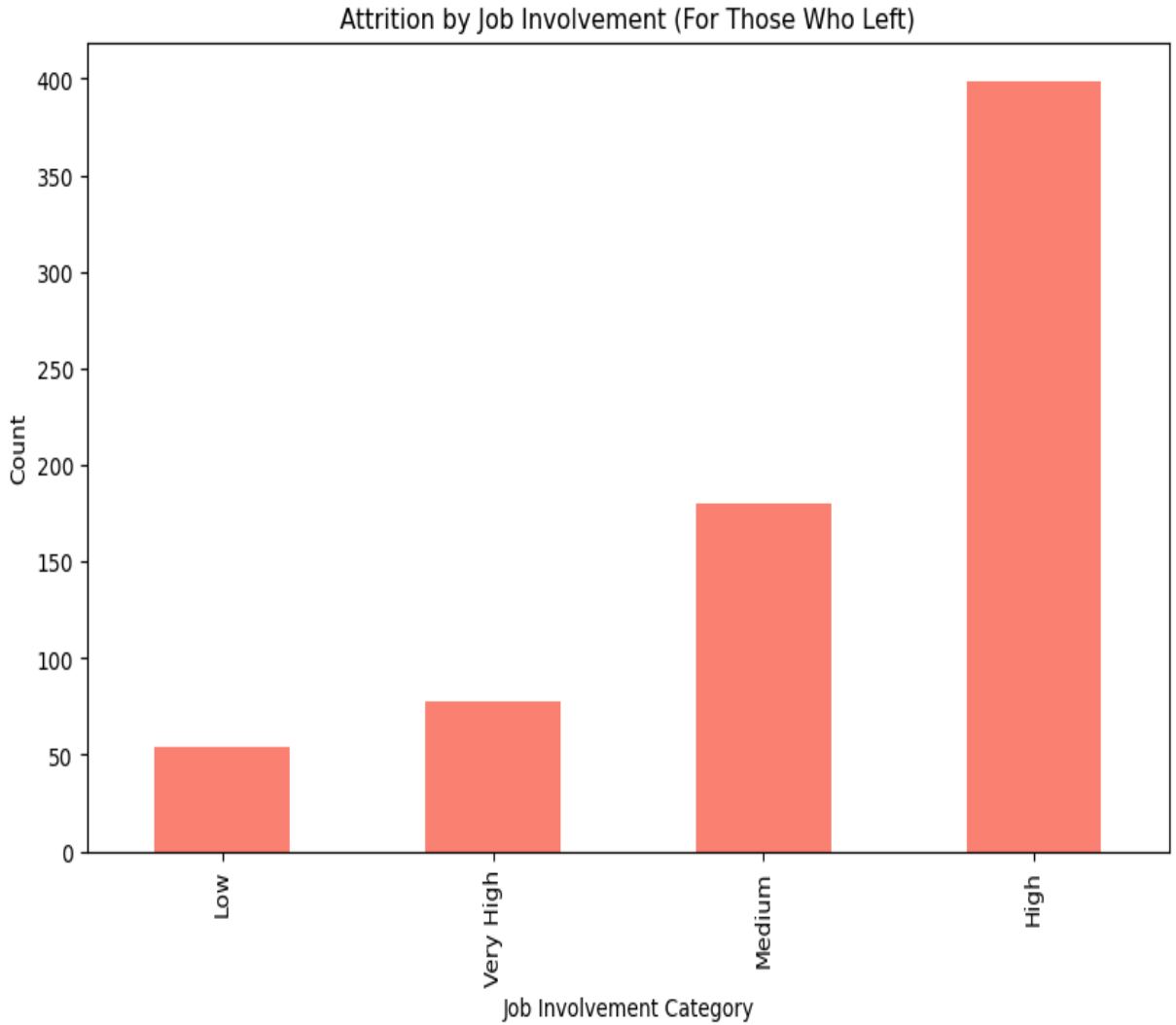
ATTRITION BY TRAVEL

- Attrition goes up with those who travel more vs no travel.



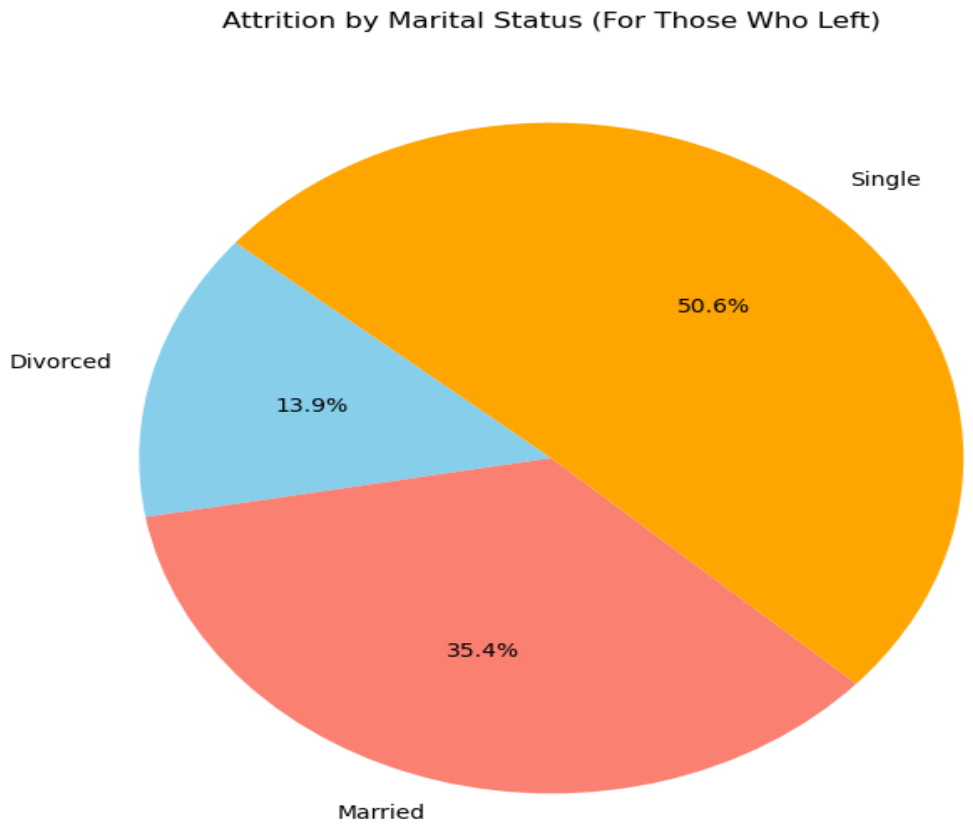
ATTRITION BY JOB EXPANSION

- Except for high job expansion- attrition count goes up based job involvement scores.



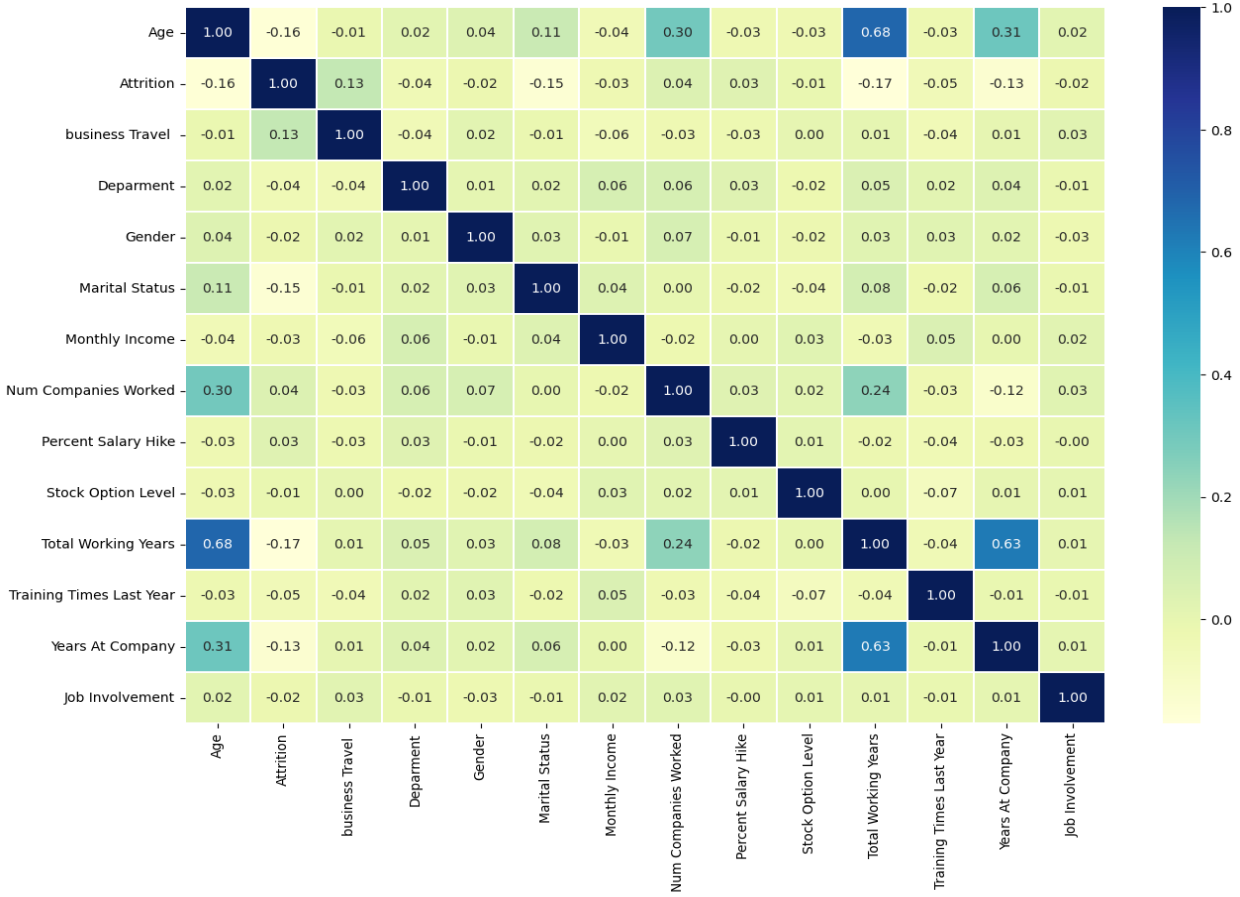
ATTRITION BY MARITAL STATUS

- Singles hold the highest attrition count amount among divorced and married.



CORRELATION CHART PRE MODEL WORK

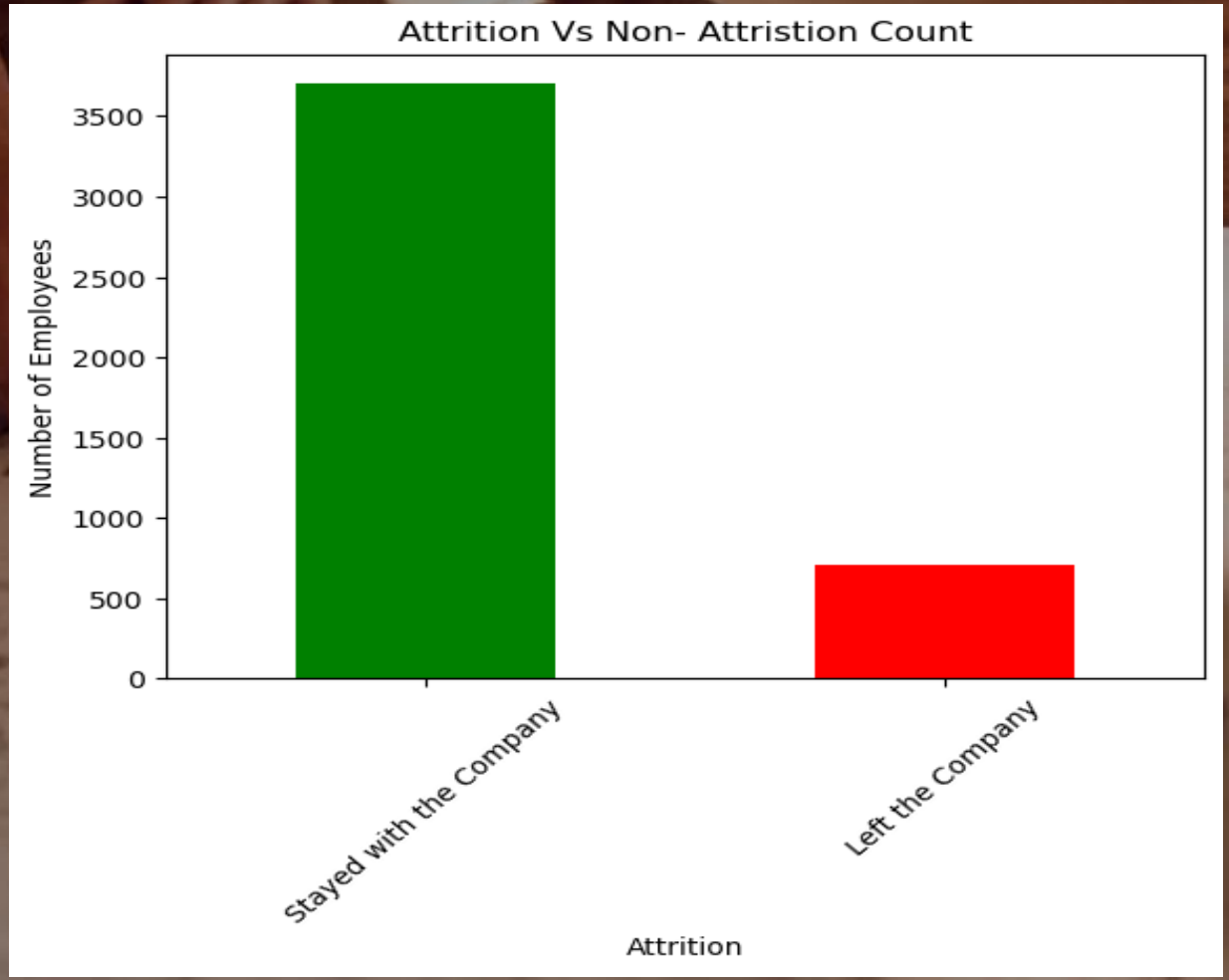
- Negative Correlation Between Attrition and Total Working Years. - 0.170338 between Attrition and Total Working Years suggests that employees who have been with the company longer are less likely to leave.
- Negative Correlation Between Attrition and Marital Status: The correlation of -0.145985




THANK YOU

ATTRITION VS NON- ATTRITION

- 3,699 Stayed from the Previous year
- 711 Employees Left from the Previous Year



1.Precision:

1. **For class 0 (no attrition)**: 0.99
2. **For class 1 (attrition)**: 1.00
3. Precision is the ratio of correctly predicted instances of a class (true positives) to all instances predicted as that class (true positives + false positives). High precision indicates that when the model predicts a class, it is very likely to be correct.

2.Recall:

1. **For class 0 (no attrition)**: 1.00
2. **For class 1 (attrition)**: 0.96
3. Recall is the ratio of correctly predicted instances of a class (true positives) to all instances of that class (true positives + false negatives). High recall indicates that the model successfully identifies most of the instances of a class.

3.F1-score:

1. **For class 0 (no attrition)**: 1.00
2. **For class 1 (attrition)**: 0.98
3. F1-score is the harmonic mean of precision and recall. It balances precision and recall and is particularly useful when the class distribution is imbalanced.

4.Support:

1. **For class 0 (no attrition)**: 741 instances
2. **For class 1 (attrition)**: 141 instances
3. Support is the number of actual instances of each class in the test set.

RECOMMEND STARTING WITH LOGISTIC REGRESS AND THEN MOVING TO RANDOM FOREST

- **Logistic Regression**:
  - **Simple and interpretable**: Logistic regression is straightforward and provides clear insights into the relationships between features and attrition.
  - **Baseline model**: It allows you to establish a baseline performance and compare it with other models.
  - **Low risk of overfitting**: Logistic regression tends to generalize well and is less prone to overfitting, making it a good starting point.
- **Random Forest**:
  - **Potential overfitting**: With high accuracy (99%), there is a possibility that the model might be too perfect and could be overfitting the training data, potentially resulting in poorer performance on unseen data.
  - **Complex model**: Random forest is a more complex model that can capture intricate patterns, but this complexity may lead to reduced interpretability and challenges in understanding its decisions.

EXECUTIVE SUMMARY

- After analyzing the accuracy scores of Logistic Regression, K-nearest neighbors, and Random Forest models, I delved deeper into Logistic Regression and Random Forest models since they had the highest accuracy scores.
- Random Forest achieved a remarkable 99 percent accuracy score, but it might be overly optimistic, potentially leading to poor performance when applied to new or unseen data.
- Specifically, Random Forest incorrectly predicted 6 instances of non-attrition (employees who stayed) as attrition,
- Logistic Regression had an accuracy of 85 percent, which is lower than Random Forest.
- Logistic Regression model demonstrates that it's not overtrained and could perform well when deployed with unseen data
- Logistic Regression only had 4 false positive predictions (incorrectly identifying employees as leaving the company)
- Had a higher number of false negative predictions, with 131 instances. This imbalance is not a significant concern since our primary focus is on attrition, not non-attrition.
- Logistic Regression showed strong alignment with the correlation matrix chart, which was consulted prior to model training. This chart provided invaluable insights into the various features and their relationship to the target variable, attrition.