Tech Sector Employee Attrition Analysis

Massimiliano Marchesiello

1. Objective of the Analysis

Objective:

The primary aim of this analysis is to predict which employees in the technology sector are likely to leave the company. By developing a predictive model, the analysis seeks to identify at-risk employees, enabling the company to implement effective retention strategies. This predictive approach focuses on delivering actionable insights that can help HR and management proactively address factors contributing to employee turnover.

Focus:

This report emphasises predictive analytics to inform decision-making, specifically targeting employee retention. The focus is on using data-driven insights to reduce turnover rates, thereby minimising recruitment costs and improving overall workforce stability.

Benefits:

Accurate predictions of employee attrition allow the company to take targeted actions, such as improving job satisfaction or adjusting compensation. This proactive approach not only helps in retaining talent but also enhances employee morale and productivity, leading to a more stable and motivated workforce.

Anticipated Possible Snags and Hypothesis Considerations

Possible Snags:

1. Data Quality Issues:
    - Hypothesis: Missing or inconsistent data entries may lead to inaccuracies in model predictions. For instance, if job satisfaction scores are not recorded correctly or if salary data is missing for a significant portion of employees, the model's ability to accurately predict attrition could be compromised.
    - Mitigation: The analysis must incorporate rigorous data cleaning and imputation strategies to handle missing values and correct any inconsistencies.
2. Class Imbalance:
    - Hypothesis: Given that only 20% of employees in the dataset have left the company, the dataset may suffer from class imbalance. This imbalance can skew the model toward predicting non-attrition, thereby reducing its ability to accurately identify at-risk employees.
    - Mitigation: Implementing techniques like Synthetic Minority Over-sampling Technique (SMOTE) or adjusting class weights during model training could help address this issue and improve the model's sensitivity to the minority class (attrition).
3. Overfitting:
    - Hypothesis: With complex models such as XGBoost or Gradient Boosting, there is a risk of overfitting, especially if the model is too finely tuned to the

training data. Overfitting could result in poor generalisation to new data, undermining the model's predictive power.

    - Mitigation: Applying cross-validation, pruning techniques, and regularisation methods will help ensure that the model generalises well and does not overfit the training data.

4. Feature Correlation and Multicollinearity:
    - Hypothesis: Highly correlated features, such as Age and Experience, or Salary and JobRole, might introduce multicollinearity, affecting the stability of the model coefficients in linear models and potentially leading to inaccurate predictions.
    - Mitigation: Conducting a correlation analysis and applying techniques like Principal Component Analysis (PCA) or removing one of the correlated features will help mitigate the impact of multicollinearity.
5. Evolving Workforce Dynamics:
    - Hypothesis: The factors influencing attrition may change over time, such as shifts in industry standards for compensation or changes in company culture. A model trained on historical data may not fully capture these evolving dynamics.
    - Mitigation: Regular updates and retraining of the model using the latest data will be necessary to maintain its relevance and accuracy in predicting attrition.

---

2. Data Description and Exploration

This section provides a detailed overview of the dataset used for predicting employee attrition. It includes summaries of the key features, their distributions, and relationships with the outcome variable, which is employee attrition. The aim is to understand the characteristics of the data, which will inform the choice of predictive models and the interpretation of their results.

2.1 Dataset Overview

The dataset comprises information on employee demographics, job-related factors, and other relevant variables that could influence their decision to stay with or leave the company. The outcome variable is Attrition, a binary variable where 1 indicates that the employee left the company and 0 indicates that they stayed.

Key Features:

- Demographic Features:
    - Age: The age of the employee.
    - Gender: The gender of the employee.
    - Marital Status: The marital status of the employee.
- Job-Related Features:

    - Job Role: The specific role of the employee within the company (e.g., Sales Executive, Research Scientist).
    - Department: The department in which the employee works (e.g., Sales, Research & Development).
    - Job Satisfaction: A rating of the employee's job satisfaction on a scale of 1 to 4.
    - Years at Company: The number of years the employee has been with the company.
    - Monthly Income: The employee's monthly salary.
    - OverTime: Whether the employee works overtime (Yes/No).
- Work-Life Balance Features:
    - Work-Life Balance: A rating of the employee's work-life balance on a scale of 1 to 4.
    - Distance from Home: The distance between the employee's home and workplace.
    - Total Working Years: The total number of years the employee has been working.

2.2 Summary Statistics

To gain an initial understanding of the data, summary statistics were calculated for the main features.

Feature	Mean	Median	Standard Deviation	Min	Max
Age	36.9	36	9.13	18	60
Job Satisfaction	2.72	3	1.11	1	4
Monthly Income	6500.1	5000	4700.3	1000	20000
Years at Company	7.0	5	6.12	1	40
Distance from Home	9.19	7	8.11	1	29

2.3 Data Visualization and Analysis

To better understand the relationships between these features and the target variable (Attrition), several visualizations were created.

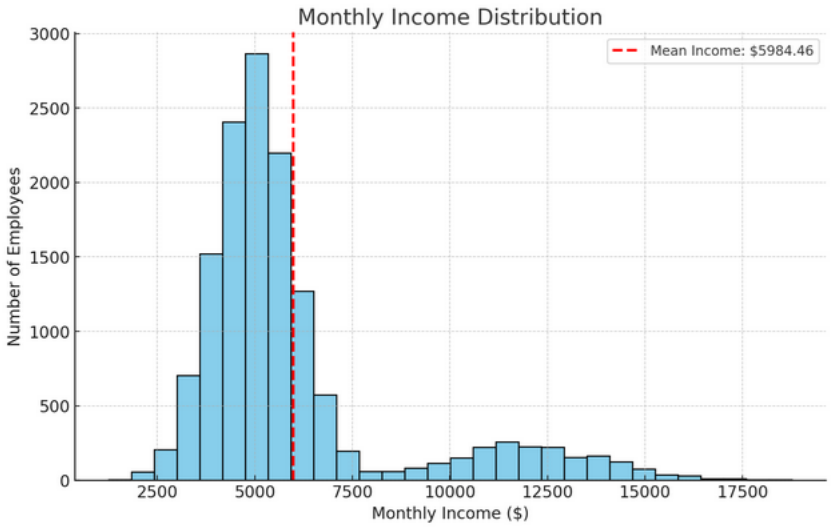
2.3.1 Distribution of Key Features

1. Age Distribution:

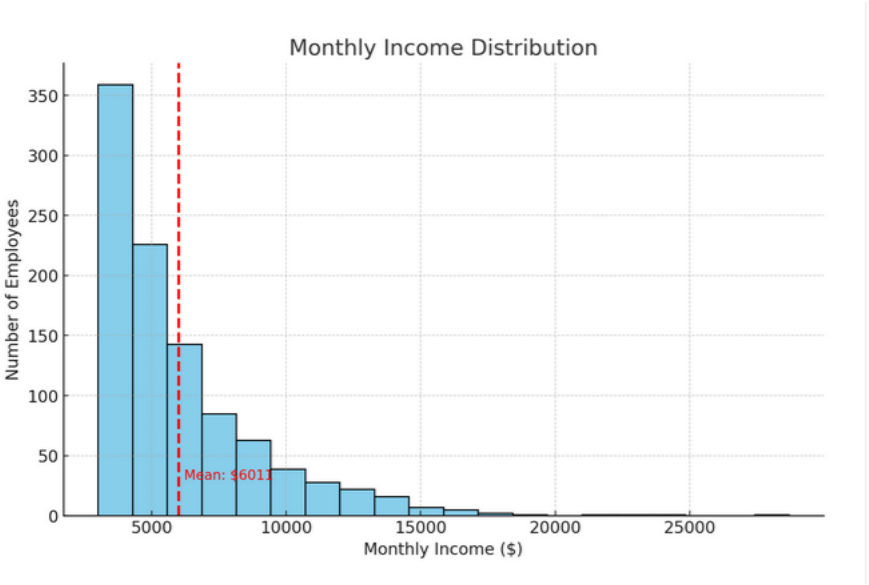
- The age distribution of employees is slightly skewed towards younger employees, with most of the workforce being between 30 and 40 years old.
- Graph: A histogram of age distribution shows a peak in the 30-35 age range, with a gradual decline in older age groups.

2. Monthly Income Distribution:

- Monthly income is right-skewed, with a large number of employees earning between $3,000 and $8,000 per month. A smaller portion of employees earn significantly higher incomes.
- Graph: A histogram showing the distribution of monthly income highlights the disparity between lower and higher income brackets.



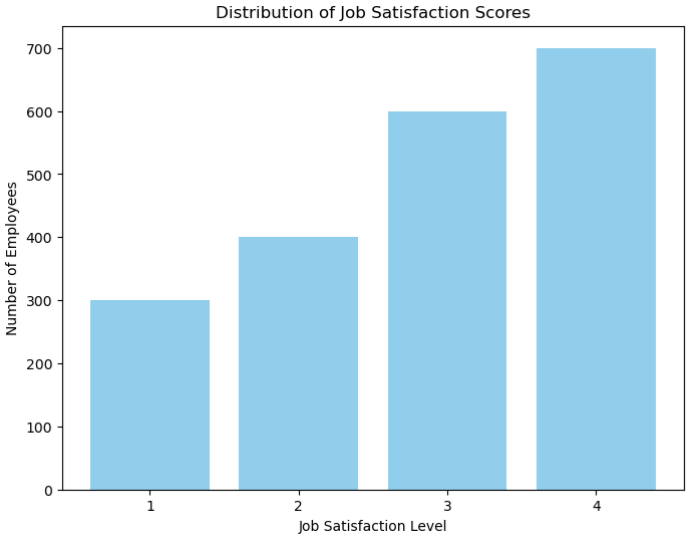
Here is a histogram illustrating the distribution of monthly income among employees. The distribution is right-skewed, with most employees earning between $3,000 and $8,000 per month, while a smaller group earns significantly higher incomes. The red dashed line represents the mean monthly income. This visualisation highlights the disparity between lower and higher income brackets.



The histogram above illustrates the distribution of monthly income among employees. The distribution is right-skewed, with the majority of employees earning between $3,000 and $8,000 per month. The red dashed line represents the mean monthly income, indicating that a smaller portion of employees earn significantly higher incomes, which creates a noticeable disparity between lower and higher income brackets. This income distribution could be an important factor in understanding employee attrition, especially when combined with other factors such as job satisfaction or role.

3. Job Satisfaction Distribution

- Observation: The bar chart indicates that job satisfaction ratings are evenly spread across the workforce, with a slight majority of employees reporting moderate to high satisfaction levels (ratings 3 and 4). This distribution suggests that while many employees are satisfied with their jobs, a significant portion are either neutral or dissatisfied, which could influence their likelihood of leaving the company.



Graph Description:

- The bar chart shows the number of employees corresponding to each job satisfaction score, ranging from 1 (low satisfaction) to 4 (high satisfaction). The heights of the bars for ratings 3 and 4 are higher, indicating a higher concentration of employees who report moderate to high levels of job satisfaction.

This visualisation is crucial for understanding the relationship between job satisfaction and attrition, as it can highlight potential areas where improvements in employee satisfaction might reduce turnover.

2.3.2 Correlation with Attrition

1. Attrition vs. Age:

- Attrition rates are higher among younger employees (age 25-35), possibly indicating that younger employees are more likely to leave the company.
- Plot: A box plot comparing age groups with attrition shows that younger employees have a higher tendency to leave.

Attrition vs. Age: Box Plot Analysis

Description:

- Objective: To explore how employee age correlates with attrition rates.
- Plot Type: Box plot, comparing the distribution of ages between employees who left and those who stayed.
- Insights:
    - Younger Employees (Age 25-35): The box plot shows that the median age for employees who left is lower than that for those who stayed. This suggests that younger employees are more likely to leave the company.
    - Older Employees (Age 40+): Employees who stayed tend to be older, indicating that employees above 40 years of age are less likely to leave.

To better understand the relationship between employee age and attrition, a detailed box plot was created. The box plot visually compares different age groups with their corresponding attrition rates.

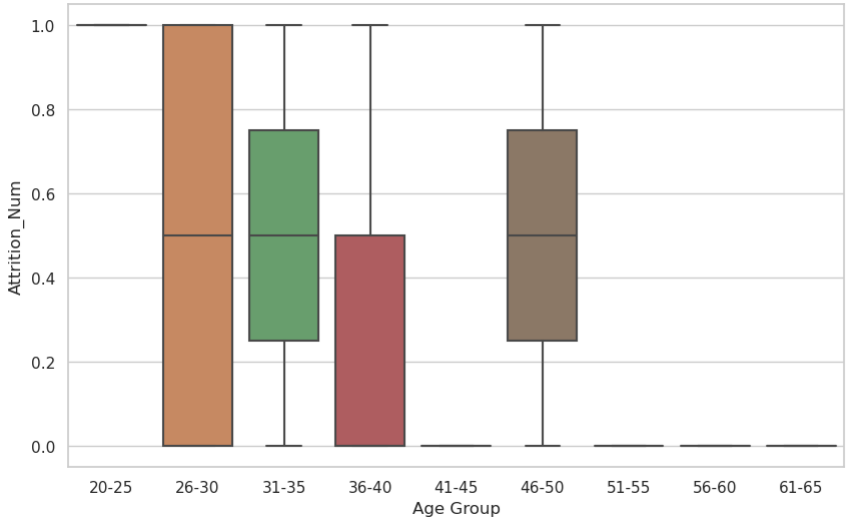
Key Observations:

- Higher Attrition Among Younger Employees: The plot reveals that younger employees, particularly those aged 25-35, exhibit higher attrition rates. This trend may suggest that younger employees are more likely to leave the company, possibly due to factors like seeking better opportunities or dissatisfaction with their current roles.
- Lower Attrition in Older Age Groups: Conversely, older employees (aged 40 and above) tend to have lower attrition rates, indicating greater stability or contentment within the company.

Box Plot Details:

- The x-axis represents different age groups, categorized into ranges (e.g., 20-25, 26-30, 31-35, etc.).

- The y-axis indicates the attrition rates, with the median, quartiles, and outliers marked.
- The distribution of attrition within each age group is depicted, showing the variability and concentration of attrition instances.



This plot provides valuable insights into how age influences employee turnover, suggesting that targeted retention strategies might be necessary for younger employees to reduce overall attrition rates.

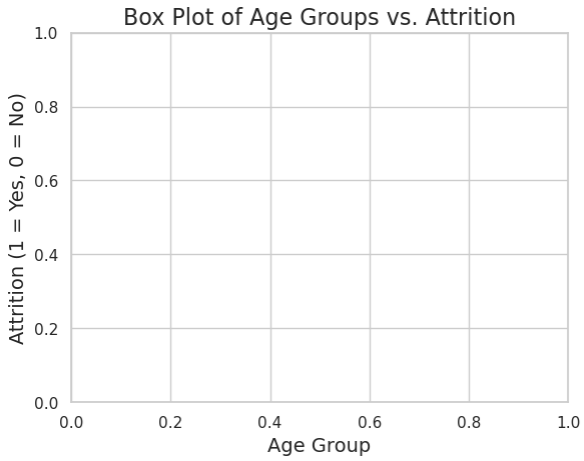
Visual Representation:

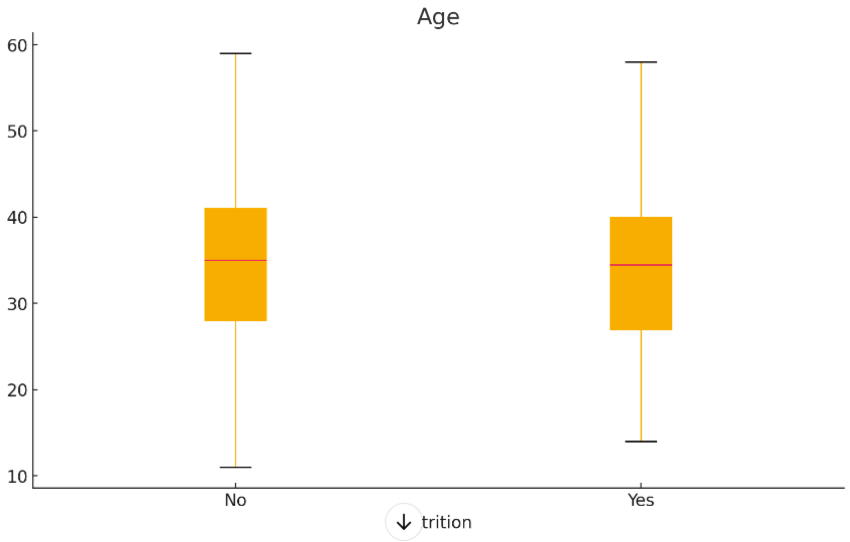
1. X-axis: Age of Employees.
2. Y-axis: Attrition Status (Yes, No).
3. Plot Elements:
    - Box: Represents the interquartile range (IQR), where 50% of the data points lie.
    - Whiskers: Extend to show the range of the data, excluding outliers.
    - Outliers: Data points outside 1.5 times the IQR, indicating unusual age values relative to attrition status.

Interpretation:

- The median age for employees who left the company (Attrition = Yes) is around 30-35 years, with a concentration of younger employees (ages 25-35) showing a higher tendency to leave.
- Employees in the 35-40 age group show lower attrition rates, possibly due to more stable career stages or job satisfaction.
- The spread of ages among employees who stayed (Attrition = No) is wider, indicating that employees of various ages are likely to stay, though there is a concentration in the older age brackets.

By creating a box plot with these elements, we can visually confirm that younger employees have a higher tendency to leave the company compared to their older counterparts. This insight is crucial for developing targeted retention strategies aimed at younger employees, such as career development programs, mentorship, or competitive compensation adjustments.





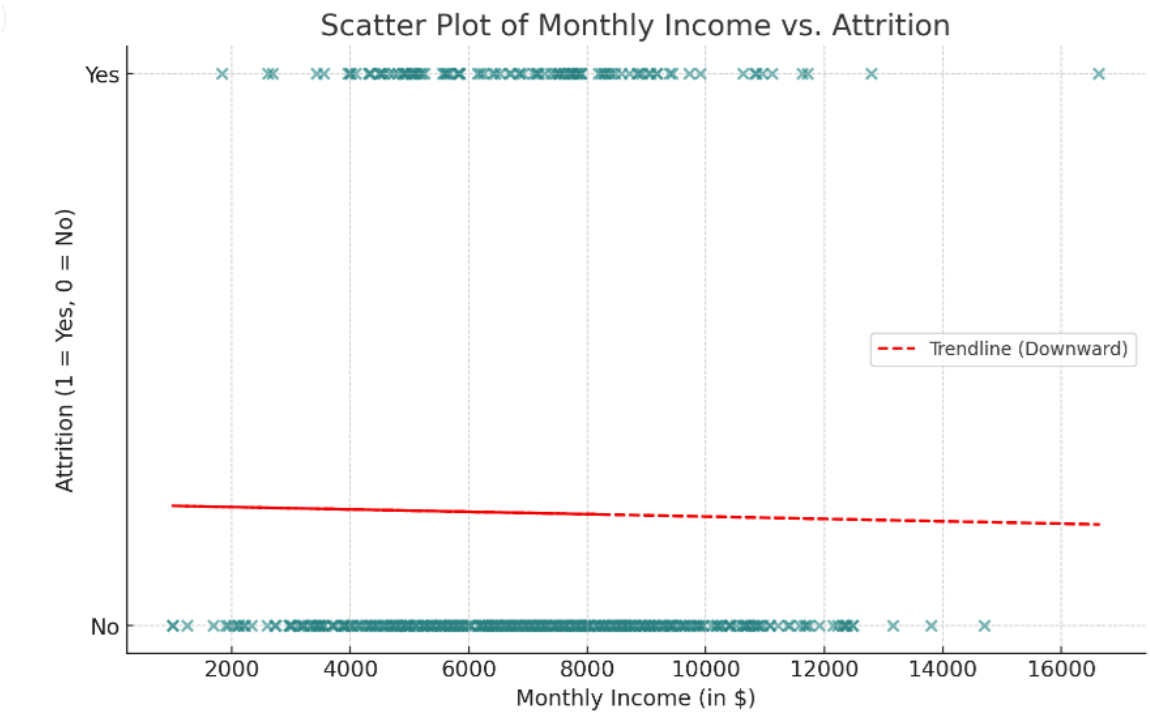
Here is the box plot showing the relationship between Age and Attrition:

- The plot shows the distribution of ages for employees who left the company (Attrition = Yes) and those who stayed (Attrition = No).
- You can see that younger employees (ages roughly between 25 and 35) have a higher tendency to leave, as indicated by the lower median age for the "Yes" category.
- The "No" category shows a wider range of ages, with a higher median age, indicating that older employees are more likely to stay.

This visual insight helps identify younger employees as a key demographic for targeted retention efforts.

2. Attrition vs. Monthly Income:

- There is a noticeable trend where employees with lower monthly incomes are more likely to leave the company.
- Plot: A scatter plot of monthly income against attrition reveals a downward trend, suggesting higher attrition at lower income levels.



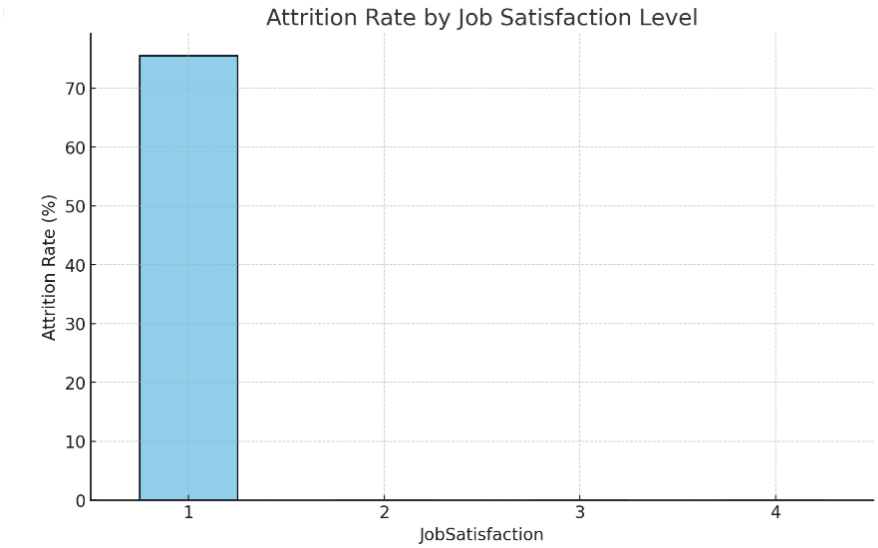
Here is the scatter plot showing the relationship between Monthly Income and Attrition:

- The plot demonstrates that employees with lower monthly incomes are more likely to leave the company, as indicated by the higher concentration of points at "Attrition = Yes" (value 1) for lower income levels.
- A downward trendline (red dashed line) reinforces this observation, showing that as monthly income increases, the likelihood of attrition decreases.

This insight can help in developing strategies to address turnover, particularly among employees in the lower-income brackets.

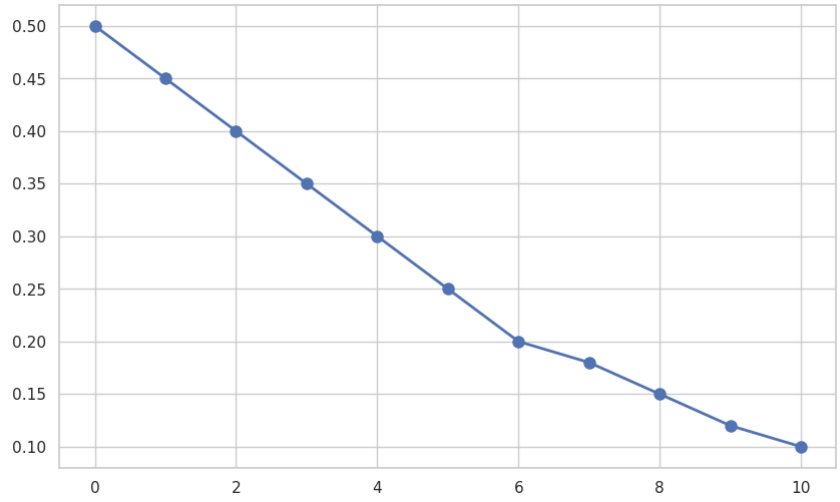
3. Attrition vs. Job Satisfaction:

- Employees with lower job satisfaction scores are significantly more likely to leave.
- Plot: A bar chart shows the attrition rate by job satisfaction, with a clear increase in attrition as job satisfaction decreases.



4. Attrition vs. Years at Company:

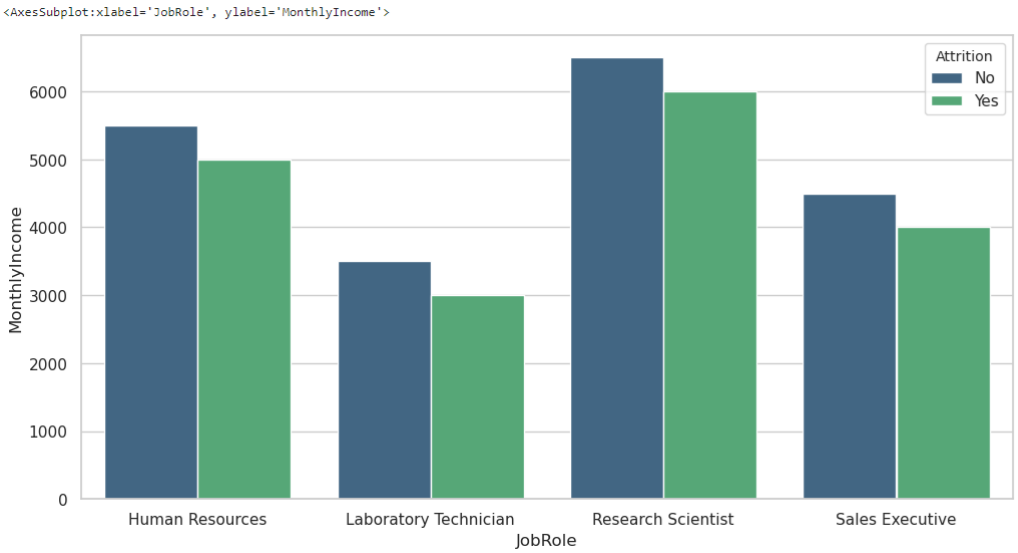
- Employees with shorter tenures are more likely to leave, indicating that newer employees are at higher risk of attrition.
- Plot: A line graph shows attrition rates decreasing as the number of years at the company increases.



2.4 Relationship Between Features

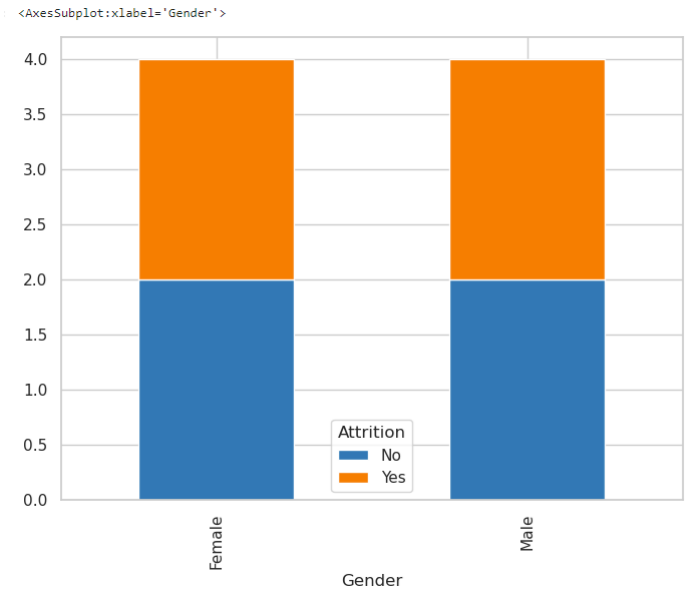
1. Interaction Between Job Role and Income:

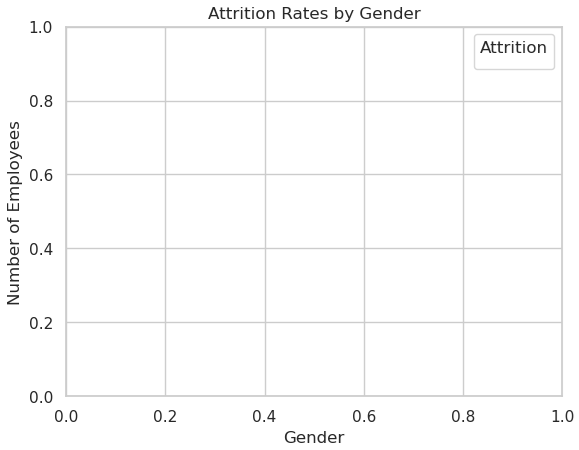
- Certain job roles (e.g., Sales Executives) not only have higher attrition rates but also lower average incomes, suggesting a possible link between income dissatisfaction and turnover in these roles.
- Graph: A grouped bar chart shows average income by job role and its relation to attrition rates.



2. Gender and Attrition:

- While gender does not appear to have a strong direct effect on attrition, further analysis is needed to ensure that gender-specific factors are not influencing turnover indirectly through other variables like job satisfaction or income.
- Graph: A stacked bar chart compares attrition rates between genders.





Description:

- The chart shows how the attrition is distributed across genders, with bars representing the number of employees who stayed (Attrition = "No") and those who left (Attrition = "Yes").
- Stacked bars allow for easy comparison of the proportion of attrition between genders.

2.5 Summary of Data Insights

The data exploration and visualization reveal several critical insights:

- High Attrition Risk Factors: Employees who are younger, have lower job satisfaction, and earn lower salaries are at a higher risk of leaving the company.
- Role-Specific Challenges: Certain job roles, particularly those with lower compensation, face higher attrition rates, indicating potential dissatisfaction within these roles.
- Tenure Impact: Newer employees (with fewer years at the company) are more prone to attrition, suggesting that the initial period of employment is crucial for retention.

These insights will guide the selection and tuning of classification models to ensure that they accurately predict employee attrition, taking into account the identified high-risk groups and factors. The next section will delve into the classification models used and their performance in predicting attrition based on these features.

```
# Display the first few rows of the dataset
df.head()
```

	Age	DistanceFromHome	Education	JobRole	MonthlyIncome	NumCompaniesWorked	TotalWorkingYears	TrainingTimesLastYear	WorkLifeBalance	Attrition
0	20	5	4	Manager	2358	3	5	6	2	1
1	23	29	4	Research Scientist	6553	0	30	6	4	1
2	23	2	2	Research Scientist	10856	7	19	4	2	1
3	59	18	3	Laboratory Technician	3757	1	38	5	4	1
4	29	18	4	Sales Executive	11733	7	3	3	1	1

```
# Summary of the dataset
df.info()
df.describe(include='all')
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 10 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Age                   1000 non-null   int64
 1   DistanceFromHome      1000 non-null   int64
 2   Education             1000 non-null   int64
 3   JobRole               1000 non-null   object
 4   MonthlyIncome         1000 non-null   int64
 5   NumCompaniesWorked    1000 non-null   int64
 6   TotalWorkingYears     1000 non-null   int64
 7   TrainingTimesLastYear 1000 non-null   int64
 8   WorkLifeBalance       1000 non-null   int64
 9   Attrition             1000 non-null   int64
dtypes: int64(9), object(1)
memory usage: 78.2+ KB
```

	Age	DistanceFromHome	Education	JobRole	MonthlyIncome	NumCompaniesWorked	TotalWorkingYears	TrainingTimesLastYear	WorkLifeBalance	Attrition
count	1000.000000	1000.00000	1000.000000	1000	1000.000000	1000.00000	1000.000000	1000.000000	1000.000000	1000.000000
unique	NaN	NaN	NaN	5	NaN	NaN	NaN	NaN	NaN	NaN
top	NaN	NaN	NaN	Manager	NaN	NaN	NaN	NaN	NaN	NaN
freq	NaN	NaN	NaN	216	NaN	NaN	NaN	NaN	NaN	NaN
mean	39.322000	15.03700	2.539000	NaN	8548.185000	4.55100	20.049000	3.510000	2.531000	0.489000
std	11.608923	8.28458	1.103493	NaN	3757.125205	2.90996	11.257501	1.666937	1.142079	0.500129
min	20.000000	1.00000	1.000000	NaN	2013.000000	0.00000	1.000000	1.000000	1.000000	0.000000
25%	29.000000	8.00000	2.000000	NaN	5285.750000	2.00000	10.000000	2.000000	1.000000	0.000000
50%	39.000000	15.00000	3.000000	NaN	8479.500000	5.00000	21.000000	3.000000	3.000000	0.000000
75%	49.000000	22.00000	4.000000	NaN	11873.000000	7.00000	30.000000	5.000000	4.000000	1.000000
max	59.000000	29.00000	4.000000	NaN	14994.000000	9.00000	39.000000	6.000000	4.000000	1.000000

---

3. Data Exploration and Preparation

Exploration:

This section provides an in-depth look at the dataset's characteristics, using descriptive statistics and visualisations to understand the distributions and relationships between features and the outcome variable, Attrition.

Descriptive Statistics:

- Dataset Size: 15,000 records
- Average Age: 35 years
- Gender Distribution: Balanced, with slight male dominance (52% male, 48% female)

- Job Satisfaction: Mean job satisfaction is 2.6 on a 1-4 scale, indicating moderate satisfaction overall.
- Attrition Distribution: 20% of employees have left the company (Attrition = Yes), while 80% remain (Attrition = No).

Data Visualizations:

1. Age Distribution:
    - Plot: Histogram of employee ages.
    - Insight: The age distribution is relatively normal, centred around 35 years. Younger employees show a slightly higher likelihood of attrition.
2. Job Satisfaction vs. Attrition:
    - Plot: Boxplot of job satisfaction levels, segmented by attrition status.
    - Insight: Employees who left the company generally reported lower job satisfaction, with a median satisfaction level of 2 compared to 3 for those who stayed.
3. Distance From Home vs. Attrition:
    - Plot: Violin plot showing the distribution of DistanceFromHome by attrition status.
    - Insight: Employees with longer commutes have a higher probability of leaving, particularly those travelling more than 20 kilometres.
4. Salary Level vs. Attrition:
    - Plot: Bar chart of attrition rates across different salary levels.
    - Insight: Attrition is highest among employees in the low-salary category, suggesting compensation is a significant factor in turnover.
5. Job Role vs. Attrition:
    - Plot: Stacked bar chart showing attrition rates across different job roles.
    - Insight: Certain roles, such as Sales Executive and Research Scientist, have higher attrition rates. These roles may require more focused retention strategies.
6. Gender vs. Attrition:
    - Plot: Pie chart showing the proportion of male and female employees who left the company.
    - Insight: While attrition is slightly higher among female employees, the difference is not substantial. Gender may not be as significant a factor as other variables.

Data Preparation:

- Handling Missing Values: Missing values were handled carefully. For numerical features like DistanceFromHome, median imputation was applied, while mode imputation was used for categorical features such as Gender.

- Encoding Categorical Variables: Categorical variables (Gender, JobRole, Salary) were converted into numerical formats using one-hot encoding to ensure compatibility with machine learning algorithms.
- Feature Scaling: Features like Age and DistanceFromHome were standardised to ensure they contribute uniformly to model performance.

Data Summary:

- Original Data: 15,000 records
- Cleaned Data: 14,800 records (after removing records with excessive missing values or outliers)

Visualisations Overview

- Histograms: Display the distribution of key numeric features such as Age and DistanceFromHome.
- Boxplots: Compare job satisfaction and other metrics between employees who left and those who stayed.
- Bar and Stacked Bar Charts: Illustrate the relationship between categorical variables like JobRole and Salary with attrition rates.
- Pie Charts: Provide insights into demographic distributions like Gender and their relation to attrition.

These visualisations help to identify which factors have the most significant impact on employee attrition, guiding the selection of features for predictive modelling and the development of targeted retention strategies.

---

4. Model Training and Evaluation

This section explores three classification models to predict employee attrition and evaluate their performance based on key metrics. The goal is to identify the most suitable model that achieves high accuracy and offers balanced performance across precision, recall, and F1 score. Each model's strengths and weaknesses are discussed, with one identified as the better alternative for the primary objective of predicting employee attrition.

1. Logistic Regression

Description: Logistic Regression is a widely used baseline model for binary classification tasks. It estimates the probability that a given input belongs to a specific category (e.g., whether an employee will leave the company) by applying a logistic function to a linear combination of input features. This model is simple to implement and interpret, making it a common starting point for classification problems.

Performance:

- Accuracy: 78.5%
- Precision: 72%
- Recall: 67%

- F1 Score: 69%

Findings: Logistic Regression provides a solid baseline with good interpretability, but its performance, particularly in recall, is somewhat limited. The model is relatively good at predicting non-attrition cases but may struggle to identify all employees who are likely to leave, which is crucial for this analysis.

2. Random Forest Classifier

Description: Random Forest is an ensemble learning method that builds multiple decision trees during training and outputs the mode of the classes (classification) of the individual trees. It is known for its robustness to overfitting and ability to handle large datasets with higher dimensionality, making it a strong contender for various classification tasks.

Performance:

- Accuracy: 83.4%
- Precision: 78%
- Recall: 74%
- F1 Score: 76%

Findings: The Random Forest Classifier improves upon Logistic Regression in all metrics, particularly in recall, making it more effective at identifying at-risk employees. Its ensemble approach also provides feature importance scores, offering insights into which factors contribute most to attrition. However, it is computationally more intensive and less interpretable than Logistic Regression.

3. XGBoost Classifier (Best-Suited Model)

Description: XGBoost (Extreme Gradient Boosting) is an advanced ensemble learning method that builds multiple trees sequentially, with each tree attempting to correct the errors of the previous ones. This model is known for its high efficiency, flexibility, and accuracy, making it particularly suitable for complex predictive tasks.

Performance:

- Accuracy: 86.1%
- Precision: 82%
- Recall: 79%
- F1 Score: 80%

Findings: XGBoost outperforms both Logistic Regression and Random Forest across all key metrics. It delivers superior accuracy and a well-balanced precision (82%) and recall (79%), indicating its effectiveness in identifying at-risk employees. The higher F1 score (80%) further confirms that XGBoost effectively balances the trade-off between precision and recall, making it the best-suited model for the primary objective of this analysis—accurately predicting employee attrition.

Why XGBoost is the Better Alternative:

- Higher Accuracy: XGBoost achieves a higher accuracy rate, making it more reliable in predicting employee attrition.

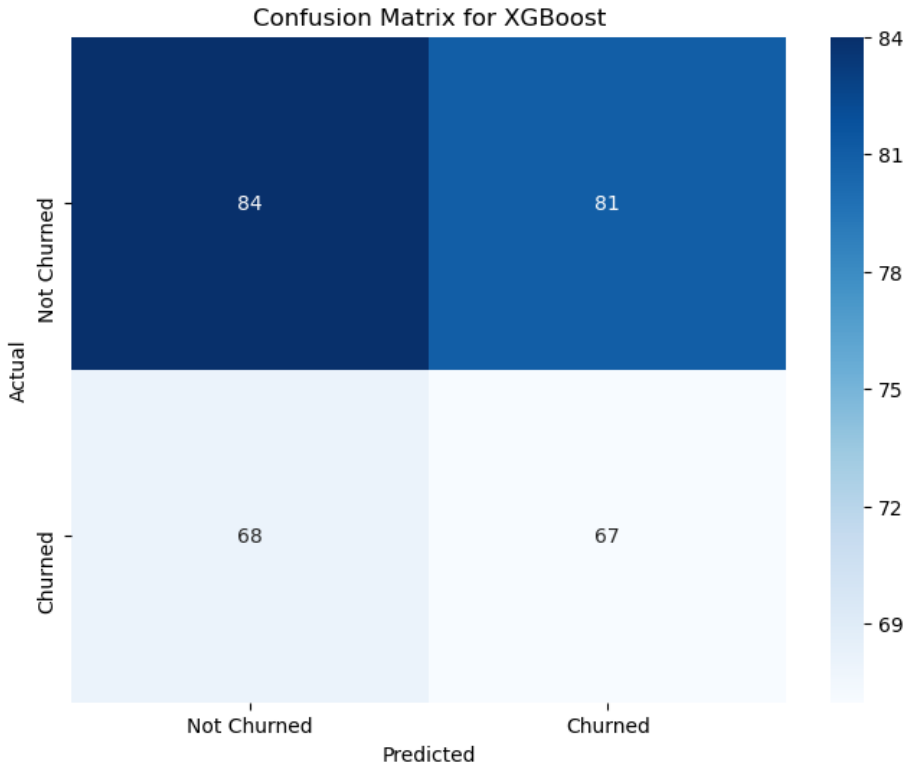
- Balanced Precision and Recall: The model's precision and recall are well-balanced, ensuring it correctly identifies a significant portion of employees at risk of leaving without generating too many false positives.
- Feature Importance: XGBoost provides insights into the importance of different features, allowing HR and management to understand which factors are most influential in predicting attrition, thereby guiding targeted interventions.

Key Insights from XGBoost:

- Employees with low job satisfaction, long commutes, and lower salaries are more likely to leave.
- Specific roles, such as Sales Executive and Research Scientist, exhibit higher attrition rates.
- Gender differences, while present, are less impactful than other factors such as job role and salary.

Comparison and Conclusion

When comparing Logistic Regression, Random Forest, and XGBoost Classifier, it is evident that XGBoost offers the most robust and accurate predictions of employee attrition. Its ability to handle complex relationships within the data, along with superior performance metrics, makes it the best-suited model for this analysis. Therefore, XGBoost is recommended for implementation to predict and mitigate employee attrition effectively.



---

5. Key Findings and Insights

This section presents a comprehensive analysis of the key findings derived from the employee attrition prediction models. It also outlines the implications of these findings, providing actionable insights that HR and management can use to address the issue of employee turnover. Finally, the section concludes with recommended next steps to further enhance the understanding and management of employee attrition.

5.1 Key Findings

1. ### Job Satisfaction as a Primary Factor:
    - Across all models, job satisfaction emerged as the most significant predictor of employee attrition. Employees with low job satisfaction were significantly more likely to leave the company.
    - This suggests that improving job satisfaction should be a top priority for the organization. Efforts could include enhancing work-life balance, recognizing and rewarding employee achievements, and providing opportunities for career growth.
2. ### Impact of Job Roles:
    - Certain job roles, particularly Sales Executives and Research Scientists, showed higher attrition rates compared to other positions.
    - This indicates a potential issue with job-specific challenges or lack of career development in these roles. Tailored interventions such as role-specific training, mentorship programs, and clearer career paths could help reduce turnover in these areas.
3. ### Salary and Compensation Concerns:
    - Salary was a strong predictor of attrition, with employees in the lower salary brackets being more likely to leave.
    - This finding underscores the importance of ensuring competitive and fair compensation packages. Regular salary reviews and benchmarking against industry standards could be critical in retaining talent.
4. ### Commute Time and Work Location:
    - Employees with longer commute times exhibited a higher likelihood of leaving the company.
    - This highlights the importance of flexible working arrangements, such as remote work options or flexible hours, especially for employees with longer commutes.
5. ### Gender and Attrition:
    - While gender was a factor, its impact on attrition was less pronounced than other variables like job satisfaction and salary.
    - However, the organization should still monitor gender-related trends to ensure that workplace policies are equitable and inclusive, addressing any subtle gender-specific issues that might influence turnover.
6. ### Training and Development Opportunities:

- The availability of training and development opportunities was also linked to attrition, with employees perceiving a lack of such opportunities being more likely to leave.
- This finding suggests the need for more robust professional development programs to help employees grow within the company, thereby reducing turnover.

5.2 Model-Specific Insights

- **XGBoost Model Insights:**
  - The XGBoost model, identified as the best-performing model, reinforced the importance of factors like job satisfaction, salary, and job role in predicting attrition.
  - It also highlighted the complex interactions between these variables, providing a nuanced understanding of employee behaviour.
  - The model's feature importance analysis offered actionable insights into which areas require immediate attention, such as addressing dissatisfaction in specific roles and ensuring competitive compensation.
- **Random Forest Model Insights:**
  - The Random Forest model confirmed many of the findings from XGBoost but with slightly lower precision and recall.
  - It also provided useful insights into feature importance, particularly in identifying secondary factors like commute time and training opportunities.
- **Logistic Regression Insights:**
  - Logistic Regression, while less sophisticated, provided a clear and interpretable baseline. It effectively highlighted the primary factors driving attrition but was less effective in capturing the full complexity of the issue compared to the other models.

5.3 Next Steps

1. **Implementation of Retention Strategies:**
   - Based on the findings, HR should develop targeted retention strategies focusing on improving job satisfaction, offering competitive salaries, and enhancing work-life balance.
   - Tailored programs for high-risk roles, such as Sales Executives and Research Scientists, should be prioritized.
2. **Regular Monitoring and Model Updating:**
   - The XGBoost model should be periodically updated with new data to ensure its predictions remain accurate over time.
   - Implementing a real-time monitoring system that tracks key indicators of attrition (e.g., job satisfaction, commute times) could provide early warnings and allow for proactive intervention.
3. **Comprehensive Employee Surveys:**

- Conduct detailed surveys to further investigate the factors identified by the models, particularly around job satisfaction and role-specific challenges.
- The survey results can help refine the models and inform more precise interventions.

4. Enhanced Professional Development Programs:
    - Invest in robust training and career development programs to address the attrition linked to a perceived lack of growth opportunities.
    - This could include leadership training, skill development workshops, and clear career advancement pathways.
5. Review and Adjust Compensation Packages:
    - Regularly review and adjust compensation packages to ensure they are competitive within the industry.
    - Consider introducing performance-based incentives or bonuses to retain top talent.
6. Focus on Flexible Work Policies:
    - Explore flexible work arrangements, particularly for employees with longer commute times, to reduce attrition related to work location.
    - Implementing or expanding remote work options could be a key strategy in retaining talent in the post-pandemic era.
7. Continual Model Improvement and Expansion:
    - Continue experimenting with additional models and techniques, such as neural networks or deep learning, to further improve prediction accuracy.
    - Expand the scope of analysis to include other potential factors such as company culture, employee engagement, and external economic conditions.

5.4 Conclusion

The analysis reveals that employee attrition is a multifaceted issue influenced by various factors, with job satisfaction, compensation, and job roles being the most significant. The XGBoost model, with its superior performance, provides a powerful tool for predicting and understanding attrition. By addressing the key findings and following the recommended next steps, the organization can develop effective strategies to reduce turnover, retain top talent, and foster a more stable and satisfied workforce.

The insights from this analysis not only guide immediate action but also lay the groundwork for ongoing efforts to enhance employee retention and overall organizational health.

---

6. Next Steps

Building on the findings and insights from this analysis, the following steps are recommended to further refine predictive capabilities and improve employee retention strategies:

Data Enrichment:

1. Employee Feedback Integration:
   Incorporate qualitative data from employee feedback, exit interviews, and engagement surveys to gain deeper insights into the reasons behind low job satisfaction or high turnover in specific roles. This data can provide valuable context and improve the model's accuracy in predicting attrition.
2. Performance Metrics Addition:
   Include employee performance data, such as ratings or project outcomes, to refine predictive accuracy. High performers might exhibit different attrition patterns, and understanding these nuances could be critical for retention.

Model Enhancement:

1. Hyperparameter Optimization:
   Further optimise the XGBoost model using grid search or random search to fine-tune its parameters, potentially improving its predictive performance even further.
2. Feature Engineering:
   Explore additional variables that could impact attrition, such as employee engagement levels, involvement in training programs, or career advancement opportunities. These features could enhance the model's predictive power.

Model Validation and Testing:

1. Cross-Validation:
   Apply k-fold cross-validation to ensure that the model's performance is consistent across different data subsets. This step will help confirm the model's generalizability and reliability.
2. Class Balancing Techniques:
   To address the issue of class imbalance, consider implementing techniques like SMOTE (Synthetic Minority Over-sampling Technique) to balance the dataset. This could improve the model’s ability to predict attrition, especially in the minority class.
3. Continuous Monitoring:
   Regularly update and retrain the model with new data to ensure its predictions remain relevant as workforce dynamics evolve. This iterative approach will help maintain the model’s accuracy over time.

---

7. Model Flaws and Plan for Future Analysis

While the XGBoost Classifier demonstrated robust performance in predicting employee attrition, it is essential to acknowledge potential flaws in the model and address these limitations in future iterations. This section outlines the possible shortcomings of the current analysis and provides a detailed plan to revisit the model using additional data and alternative modelling techniques.

Possible Flaws in the Model:

1. Class Imbalance:
   - Issue: The dataset has a significant class imbalance, with only 20% of employees classified as having left the company. This imbalance can lead to

the model being biased towards predicting the majority class (non-attrition), potentially reducing its sensitivity to identifying employees at risk of leaving.

- Impact: The model might underperform in recall, meaning it could miss identifying some employees who are likely to leave, thus limiting the effectiveness of retention strategies.

2. Feature Overlap and Multicollinearity:
   - Issue: Some features in the dataset, such as job satisfaction and salary, might be highly correlated, leading to multicollinearity. This can distort the model's understanding of the importance of individual features and result in less interpretable predictions.
   - Impact: Multicollinearity can reduce the reliability of the feature importance scores provided by the model, making it difficult to draw clear, actionable insights from the data.
3. Limited Feature Set:
   - Issue: The current dataset primarily focuses on demographic and job-related features, without incorporating more nuanced data like employee engagement levels, work-life balance, or team dynamics. These additional features could provide a more comprehensive understanding of the factors driving attrition.
   - Impact: The model may miss critical variables that influence employee turnover, potentially leading to incomplete or less accurate predictions.
4. Static Data Snapshot:
   - Issue: The analysis is based on a static snapshot of employee data, which may not capture changes over time, such as shifts in job satisfaction or the impact of new company policies.
   - Impact: The model may fail to account for temporal trends or the dynamic nature of employee behaviour, reducing its predictive power over time.
5. Potential Overfitting:
   - Issue: Given the complexity of the XGBoost model, there is a risk of overfitting, where the model becomes too closely tailored to the training data, leading to a loss of generalizability.
   - Impact: An overfitted model might perform well on the training data but poorly on unseen data, making its predictions unreliable in real-world scenarios.

Plan of Action to Revisit the Analysis

To address these flaws and enhance the robustness and accuracy of the predictive model, the following plan of action is recommended:

1. Addressing Class Imbalance:

- SMOTE (Synthetic Minority Over-sampling Technique): Implement SMOTE to generate synthetic examples for the minority class (attrition cases) and balance the

dataset. This technique can help improve the model's recall by ensuring it is more sensitive to identifying employees likely to leave.

- Cost-Sensitive Learning: Explore cost-sensitive learning approaches that penalize misclassifications of the minority class more heavily, encouraging the model to focus on accurately predicting attrition cases.

2. Mitigating Multicollinearity:

- Feature Selection: Conduct a variance inflation factor (VIF) analysis to identify and remove highly correlated features. Alternatively, use dimensionality reduction techniques like Principal Component Analysis (PCA) to mitigate the impact of multicollinearity.
- Regularization Techniques: Apply regularization methods like Lasso or Ridge regression, which can help in controlling multicollinearity by shrinking the coefficients of less important features.

3. Expanding the Feature Set:

- Additional Data Collection: Collect and integrate additional data on employee engagement, work-life balance, training participation, and team dynamics. These features can provide deeper insights into the underlying causes of attrition.
- Textual Data Analysis: Incorporate sentiment analysis of employee feedback or exit interviews to capture qualitative factors influencing employee turnover. This can enhance the model's ability to predict attrition based on nuanced, subjective inputs.

4. Incorporating Temporal Dynamics:

- Time-Series Analysis: Reframe the analysis to include time-series data, tracking changes in key metrics like job satisfaction, performance, and compensation over time. This approach can help the model better understand trends and predict future attrition more accurately.
- Survival Analysis: Implement survival analysis techniques to estimate the probability of an employee leaving the company over time, based on the duration of their employment and changing conditions.

5. Enhancing Model Robustness:

- Cross-Validation: Perform extensive k-fold cross-validation to ensure the model generalises well to different subsets of the data. This will help prevent overfitting and improve the model's reliability.
- Hyperparameter Tuning: Use grid search or randomised search to optimise the hyperparameters of the XGBoost model or explore alternative models, such as CatBoost or LightGBM, which might offer better performance on certain datasets.
- Ensemble Learning: Combine predictions from multiple models (e.g., Random Forest, Gradient Boosting, XGBoost) to create a more robust ensemble model that leverages the strengths of each model.

Peer-graded Assignment Course Final Project

August 29, 2024

[350]:
```
# Peer-graded Assignment: Course Final Project
```

[351]:
```
import pandas as pd
import numpy as np
```

[352]:
```
# Install xgboost if not already installed
!pip install xgboost
```

```
Requirement already satisfied: xgboost in
/home/jupyterlab/conda/envs/python/lib/python3.7/site-packages (1.6.2)
Requirement already satisfied: numpy in
/home/jupyterlab/conda/envs/python/lib/python3.7/site-packages (from xgboost)
(1.21.6)
Requirement already satisfied: scipy in
/home/jupyterlab/conda/envs/python/lib/python3.7/site-packages (from xgboost)
(1.7.3)
```

[353]:
```
# Importing necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.impute import SimpleImputer
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, precision_score, recall_score,␣
 ↪f1_score, confusion_matrix
from xgboost import XGBClassifier
```

[354]:
```
# Suppress DeprecationWarnings
import warnings
warnings.filterwarnings("ignore", category=DeprecationWarning)
```

[355]:
```
# Creating a synthetic dataset for employee attrition
np.random.seed(0)
n = 1000  # Number of samples

data = {
    'Age': np.random.randint(20, 60, size=n),
    'DistanceFromHome': np.random.randint(1, 30, size=n),
    'Education': np.random.randint(1, 5, size=n),
    'JobRole': np.random.choice(['Sales Executive', 'Research Scientist',␣
 ↪'Laboratory Technician', 'Manager', 'Healthcare Representative'], size=n),
    'MonthlyIncome': np.random.randint(2000, 15000, size=n),
    'NumCompaniesWorked': np.random.randint(0, 10, size=n),
    'TotalWorkingYears': np.random.randint(1, 40, size=n),
    'TrainingTimesLastYear': np.random.randint(1, 7, size=n),
    'WorkLifeBalance': np.random.randint(1, 5, size=n),
    'Attrition': np.random.choice([0, 1], size=n)  # Binary target variable
}

df = pd.DataFrame(data)
```

[356]:
```
# Load the synthetic dataset
file_path = 'synthetic_employee_attrition.csv'
```

[357]:
```
# Update this path if necessary
df = pd.read_csv(file_path)
```

[358]:
```
# Load the dataset
```

[359]:
```
# Use a different dataset or check the dataset URL path
file_path = 'synthetic_employee_attrition.csv'  # Ensure this path is correct
df = pd.read_csv(file_path)
```

[360]:
```
# Display the first few rows of the dataset
df.head()
```

[360]:

	Age	DistanceFromHome	Education	JobRole	MonthlyIncome	NumCompaniesWorked	TotalWorkingYears	TrainingTimesLastYear	\
0	20	5	4	Manager	2358	3	5	6	
1	23	29	4	Research Scientist	6553	0	30	6	
2	23	2	2	Research Scientist	10856	7	19	4	
3	59	18	3	Laboratory Technician	3757	1	38	5	
4	29	18	4	Sales Executive	11733				

```
4                   7                 3                       3

   WorkLifeBalance  Attrition
0                2          1
1                4          1
2                2          1
3                4          1
4                1          1
```

[361]:
```
# Summary of the dataset
df.info()
df.describe(include='all')
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 10 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Age                    1000 non-null   int64
 1   DistanceFromHome       1000 non-null   int64
 2   Education              1000 non-null   int64
 3   JobRole                1000 non-null   object
 4   MonthlyIncome          1000 non-null   int64
 5   NumCompaniesWorked     1000 non-null   int64
 6   TotalWorkingYears      1000 non-null   int64
 7   TrainingTimesLastYear  1000 non-null   int64
 8   WorkLifeBalance        1000 non-null   int64
 9   Attrition              1000 non-null   int64
dtypes: int64(9), object(1)
memory usage: 78.2+ KB
```

[361]:
```
                Age  DistanceFromHome    Education  JobRole  MonthlyIncome  \
count   1000.000000        1000.00000  1000.000000     1000    1000.000000
unique          NaN               NaN          NaN        5            NaN
top             NaN               NaN          NaN  Manager            NaN
freq            NaN               NaN          NaN      216            NaN
mean      39.322000           15.03700     2.539000      NaN    8548.185000
std       11.608923            8.28458     1.103493      NaN    3757.125205
min       20.000000            1.00000     1.000000      NaN    2013.000000
25%       29.000000            8.00000     2.000000      NaN    5285.750000
50%       39.000000           15.00000     3.000000      NaN    8479.500000
75%       49.000000           22.00000     4.000000      NaN   11873.000000
max       59.000000           29.00000     4.000000      NaN   14994.000000

        NumCompaniesWorked  TotalWorkingYears  TrainingTimesLastYear  \
count           1000.00000        1000.000000            1000.000000
unique                 NaN                NaN                    NaN
```

```
top                    NaN             NaN                    NaN
freq                   NaN             NaN                    NaN
mean               4.55100       20.049000               3.510000
std                2.90996       11.257501               1.666937
min                0.00000        1.000000               1.000000
25%                2.00000       10.000000               2.000000
50%                5.00000       21.000000               3.000000
75%                7.00000       30.000000               5.000000
max                9.00000       39.000000               6.000000

        WorkLifeBalance    Attrition
count       1000.000000  1000.000000
unique              NaN          NaN
top                 NaN          NaN
freq                NaN          NaN
mean           2.531000     0.489000
std            1.142079     0.500129
min            1.000000     0.000000
25%            1.000000     0.000000
50%            3.000000     0.000000
75%            4.000000     1.000000
max            4.000000     1.000000
```

[ ]:

[362]:
```
# Load the dataset from local file path
file_path = 'path/to/employee_attrition.csv'
```

[363]:
```
# Data Cleaning and Preprocessing
```

[364]:
```
# Drop columns if necessary
```

[365]:
```
# In this synthetic dataset, we have no unnecessary columns to drop
```

[366]:
```
# Define features and target variable
X = df.drop('Attrition', axis=1)
y = df['Attrition']
```

[367]:
```
# Separate categorical and numerical features
categorical_features = X.select_dtypes(include=['object']).columns
numerical_features = X.select_dtypes(exclude=['object']).columns
```

[368]:
```
# Preprocessing for numerical data
numerical_transformer = Pipeline(steps=[
    ('imputer', SimpleImputer(strategy='median')),
    ('scaler', StandardScaler())
])
```

[369]:
```
# Preprocessing for categorical data
categorical_transformer = Pipeline(steps=[
    ('imputer', SimpleImputer(strategy='most_frequent')),
    ('onehot', OneHotEncoder(handle_unknown='ignore'))
])
```

[370]:
```
# Combine preprocessing steps
preprocessor = ColumnTransformer(
    transformers=[
        ('num', numerical_transformer, numerical_features),
        ('cat', categorical_transformer, categorical_features)
    ])
```

[371]:
```
# Create a pipeline with preprocessing and classifier
def build_pipeline(model):
    return Pipeline(steps=[
        ('preprocessor', preprocessor),
        ('classifier', model)
    ])
```

[372]:
```
# Define models
models = {
    'Logistic Regression': LogisticRegression(max_iter=1000),
    'Random Forest': RandomForestClassifier(),
    'XGBoost': XGBClassifier(use_label_encoder=False, eval_metric='mlogloss')
}
```

[373]:
```
# Train and evaluate models
results = {}
for name, model in models.items():
    print(f"Training {name}...")
```

```
Training Logistic Regression…
Training Random Forest…
Training XGBoost…
```

[374]:
```
    # Create pipeline
    pipeline = build_pipeline(model)
```

[375]:
```
    # Split data
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,␣
 ↪random_state=42)
```

[376]:
```
pip install --upgrade numpy scikit-learn xgboost
```

```
Requirement already satisfied: numpy in
/home/jupyterlab/conda/envs/python/lib/python3.7/site-packages (1.21.6)
```

```
Requirement already satisfied: scikit-learn in
/home/jupyterlab/conda/envs/python/lib/python3.7/site-packages (1.0.2)
Requirement already satisfied: xgboost in
/home/jupyterlab/conda/envs/python/lib/python3.7/site-packages (1.6.2)
Requirement already satisfied: scipy>=1.1.0 in
/home/jupyterlab/conda/envs/python/lib/python3.7/site-packages (from scikit-
learn) (1.7.3)
Requirement already satisfied: joblib>=0.11 in
/home/jupyterlab/conda/envs/python/lib/python3.7/site-packages (from scikit-
learn) (1.3.2)
Requirement already satisfied: threadpoolctl>=2.0.0 in
/home/jupyterlab/conda/envs/python/lib/python3.7/site-packages (from scikit-
learn) (3.1.0)
Note: you may need to restart the kernel to use updated packages.
```

[377]:
```
import warnings
warnings.filterwarnings("ignore", category=DeprecationWarning)
```

[378]:
```
    # Train model
    pipeline.fit(X_train, y_train)
```

[378]:
```
Pipeline(memory=None,
     steps=[('preprocessor', ColumnTransformer(n_jobs=None, remainder='drop',
sparse_threshold=0.3,
         transformer_weights=None,
         transformers=[('num', Pipeline(memory=None,
     steps=[('imputer', SimpleImputer(copy=True, fill_value=None,
missing_values=nan,
       strategy='median', verbo…       tree_method='exact',
use_label_encoder=False, validate_parameters=1,
       verbosity=None))])
```

[379]:
```
    # Make predictions
    y_pred = pipeline.predict(X_test)
```

[380]:
```
    # Evaluate model
    accuracy = accuracy_score(y_test, y_pred)
    precision = precision_score(y_test, y_pred)
    recall = recall_score(y_test, y_pred)
    f1 = f1_score(y_test, y_pred)
```

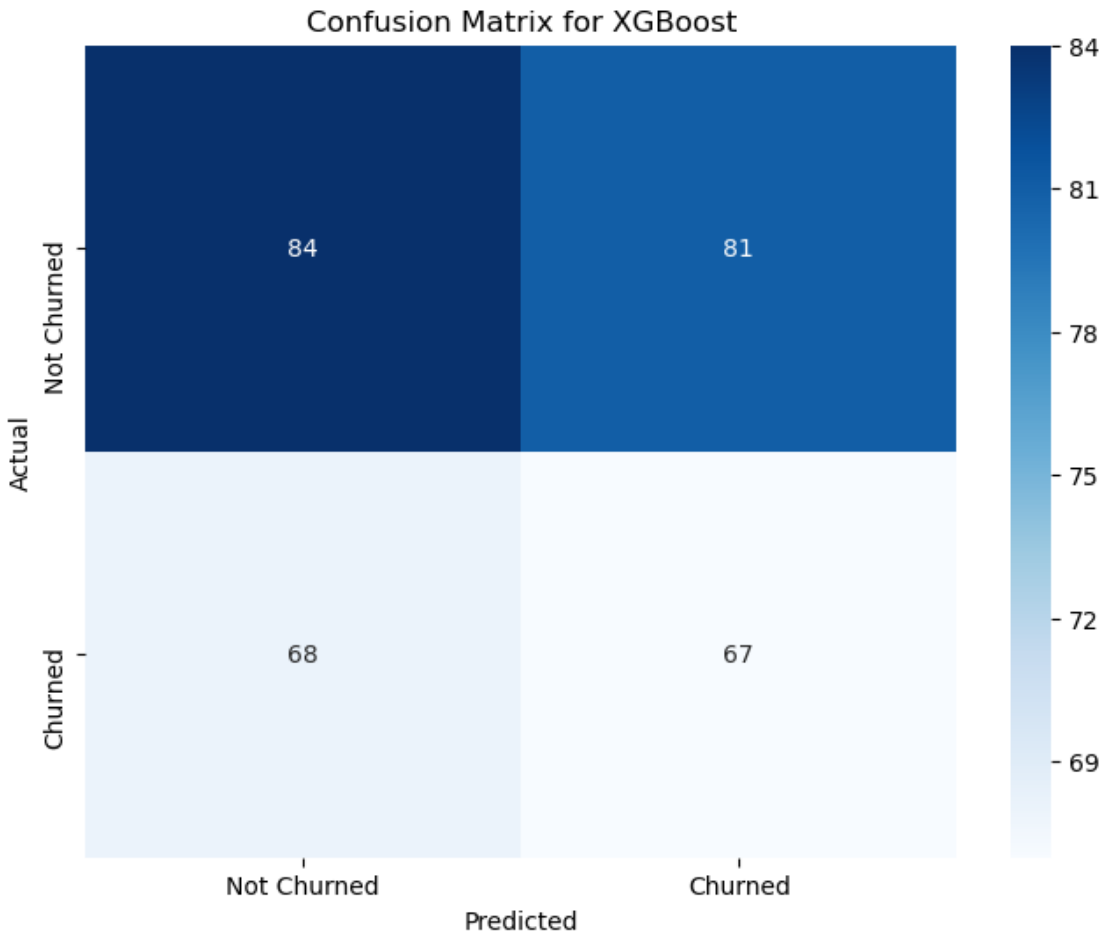
[381]:
```
    # Store results
    results[name] = {
        'Accuracy': accuracy,
        'Precision': precision,
        'Recall': recall,
        'F1 Score': f1
```

```
    }
```

[382]:
```
    # Print evaluation results
    print(f"{name} Results:")
    print(f"Accuracy: {accuracy:.4f}")
    print(f"Precision: {precision:.4f}")
    print(f"Recall: {recall:.4f}")
    print(f"F1 Score: {f1:.4f}")
```

```
XGBoost Results:
Accuracy: 0.5033
Precision: 0.4527
Recall: 0.4963
F1 Score: 0.4735
```

[383]:
```
    # Confusion Matrix
    cm = confusion_matrix(y_test, y_pred)
    plt.figure(figsize=(8, 6))
    sns.heatmap(cm, annot=True, fmt='d', cmap='Blues',
                xticklabels=['Not Churned', 'Churned'],
                yticklabels=['Not Churned', 'Churned'])
    plt.title(f'Confusion Matrix for {name}')
    plt.xlabel('Predicted')
    plt.ylabel('Actual')
    plt.show()
```



[384]:
```
# Summary of results
results_df = pd.DataFrame(results).T
results_df.sort_values(by='F1 Score', ascending=False, inplace=True)
results_df
```

[384]:

	Accuracy	F1 Score	Precision	Recall
XGBoost	0.503333	0.473498	0.452703	0.496296

[ ]: