DS 5220: Supervised Machine Learning and Learning Theory Summer, 2024

In Class Work #4

June 02, 2024

In-Class Assignment: Employee Retention Analysis and Classification

Dataset

Objective:

Analyze the employee retention dataset to identify key factors influencing employee retention and develop a logistic regression model to predict employee retention. This exercise will enhance your data analysis, visualization, and machine learning skills.

Instructions

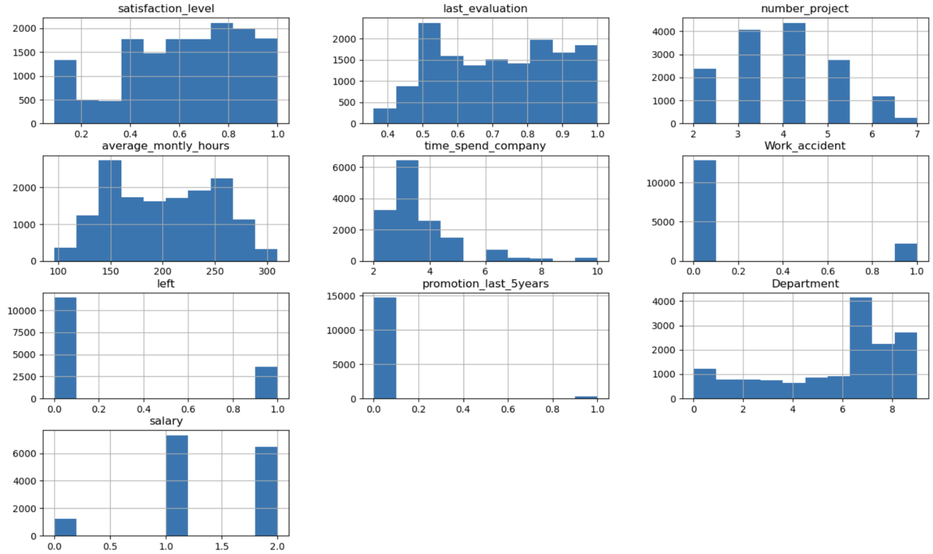
Dataset Download:

Download the employee retention dataset from Kaggle
https://www.kaggle.com/datasets/giripujar/hr-analytics

Exploratory Data Analysis (EDA):

- **Perform initial data exploration to understand the dataset structure, missing values, and basic statistics.**

  The dataset consists of 15,000 rows and 10 columns. Each column includes variables such as satisfaction level, last evaluation, number of projects, average monthly hours, time spent in the company, work accidents, employment status, promotions in the last 5 years, department, and salary. The dataset does not contain any missing values.



This graph displays histograms for each feature in the dataset, giving a visual summary of the value distribution for each variable.

1. satisfaction level: The distribution reveals a peak between 0.7 and 0.9, indicating that many employees have high satisfaction levels. There are also significant counts at lower satisfaction levels, around 0.1 to 0.2.

2. last evaluation: The last evaluation scores are fairly evenly distributed, with a peak around 0.5 to 0.6. This suggests a wide range of performance evaluation scores among employees.

3. number project: Most employees are involved in 2 to 7 projects, with a peak at 3 to 4 projects. This indicates that the majority of employees handle 3 or 4 projects.

4. average montly hours: The average monthly hours worked by employees vary widely, with noticeable peaks around 150-200 hours and 250 hours. This suggests the presence of two groups of employees, one working around 150-200 hours and another around 250 hours per month.

5. time spend company: The majority of employees have been with

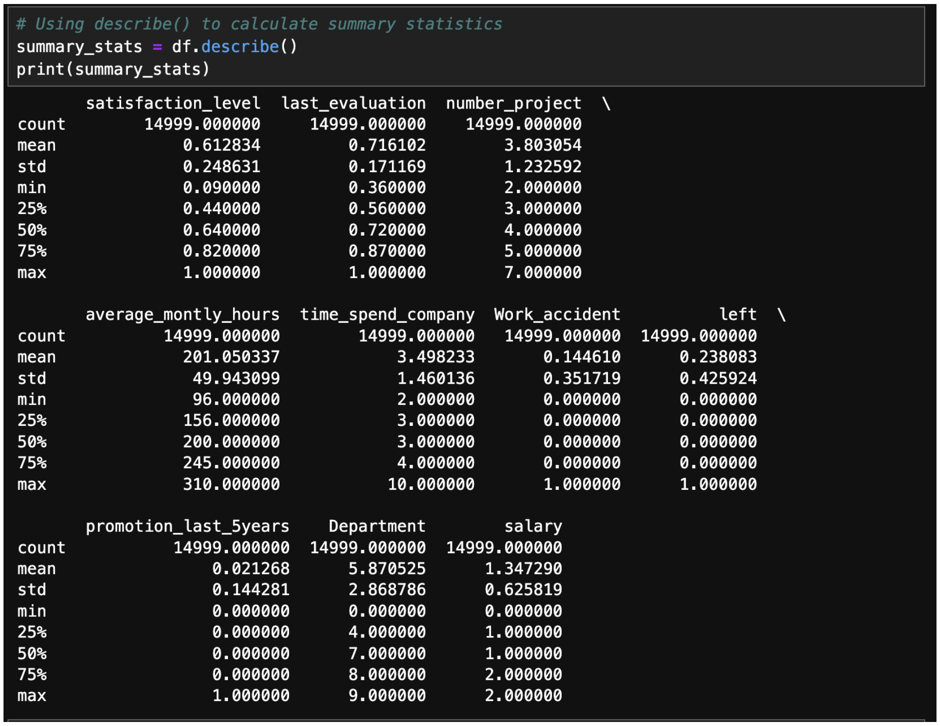
the company for 3 years, with a noticeable drop-off after that. Only a few employees stay beyond 6 years.

6. Work accident: Most employees did not experience a work accident, as indicated by the high count at 0.
7. left: This indicates a significant class imbalance, with most employees not leaving the company (0) compared to those who did (1).
8. promotion last 5years: Very few employees received a promotion in the last 5 years, as shown by the high count at 0.
9. Department: The distribution shows the number of employees in each department. Departments like sales (encoded as 7) and technical (encoded as 9) have higher counts, while others like HR (encoded as 3) and accounting (encoded as 2) have fewer employees.
10. salary: The distribution of salary levels is in three distinct groups, with medium (encoded as 1) and low (encoded as 0) salary levels having the highest counts, and high (encoded as 2) having the least.

- **Identify and discuss which variables might have a direct and clear impact on employee retention (i.e., whether they leave the company or continue to work).**

  After analyzing the data, it is unclear which variables have a direct impact on employee retention. So, further analysis will be carried out in the visualization section. By studying the graph, we can pinpoint the factors contributing to retention.

- **Use summary statistics and visualizations to support your analysis.**



```
# Using describe() to calculate summary statistics
summary_stats = df.describe()
print(summary_stats)
```

```
       satisfaction_level  last_evaluation  number_project  \
count        14999.000000     14999.000000    14999.000000   
mean             0.612834         0.716102        3.803054   
std              0.248631         0.171169        1.232592   
min              0.090000         0.360000        2.000000   
25%              0.440000         0.560000        3.000000   
50%              0.640000         0.720000        4.000000   
75%              0.820000         0.870000        5.000000   
max              1.000000         1.000000        7.000000   

       average_montly_hours  time_spend_company  Work_accident          left  \
count          14999.000000        14999.000000   14999.000000  14999.000000   
mean             201.050337            3.498233       0.144610      0.238083   
std               49.943099            1.460136       0.351719      0.425924   
min               96.000000            2.000000       0.000000      0.000000   
25%              156.000000            3.000000       0.000000      0.000000   
50%              200.000000            3.000000       0.000000      0.000000   
75%              245.000000            4.000000       0.000000      0.000000   
max              310.000000           10.000000       1.000000      1.000000   

       promotion_last_5years    Department        salary  
count           14999.000000  14999.000000  14999.000000  
mean                0.021268      5.870525      1.347290  
std                 0.144281      2.868786      0.625819  
min                 0.000000      0.000000      0.000000  
25%                 0.000000      4.000000      1.000000  
50%                 0.000000      7.000000      1.000000  
75%                 0.000000      8.000000      2.000000  
max                 1.000000      9.000000      2.000000  
```

The describe() function is used on a DataFrame that contains a dataset. This function calculates different descriptive statistics for each column in the dataset.

Here are the descriptive statistics provided by the output:

– Count: This shows the number of non-null values in each column.

– Mean: This represents the average value of the column.

– Standard Deviation (std): It measures how spread out the values are.

– Minimum (min): This is the smallest value in the column.

– 25th Percentile 25%: This is the value below which 25%of the data falls.

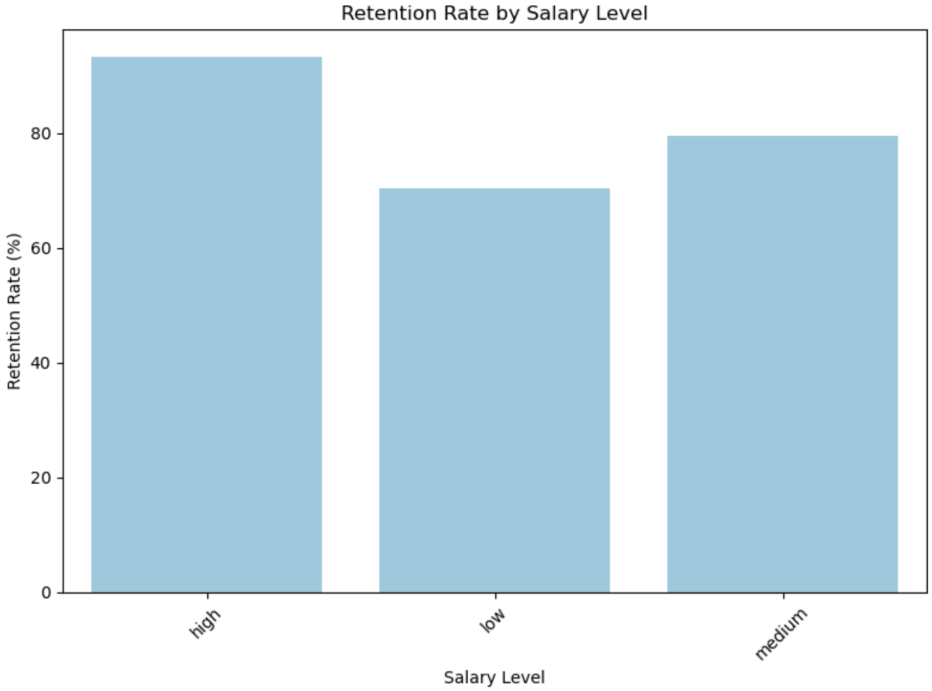
– Median 50%: This is the middle value, also known as the 50th percentile.

– 75th Percentile 75%: This is the value below which 75% of the data falls.

- Maximum (max): This is the largest value in the column.

Data Visualization:

- **Plot bar charts to show the impact of employee salaries on retention.**

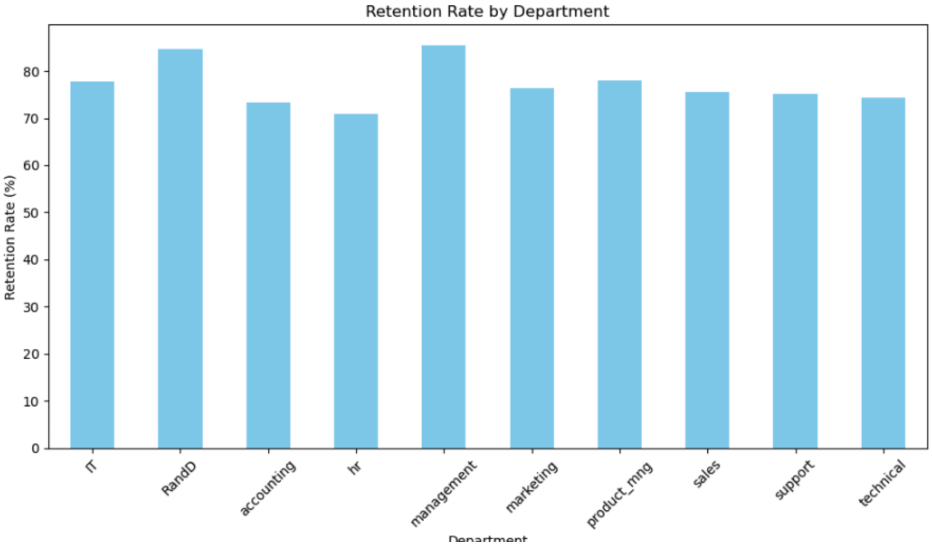


According to the bar graph titled "Retention Rate by Salary Level," we can observe the following:

- **High Salary Level:** The bar exceeds 80%, indicating that employees with higher salaries generally have higher retention rates.
- **Low Salary Level:** This bar is slightly above 60%, suggesting that employees with lower salaries may experience lower retention rates.
- **Medium Salary Level:**The bar hovers around 70%, indicating that employees with medium salaries have retention rates between those with high and low salaries.

Overall, it appears that there is a potential correlation between salary levels and employee retention rates. It implies that higher salaries could potentially contribute to higher retention rates.

- **Plot bar charts to show the correlation between departments and employee retention.**



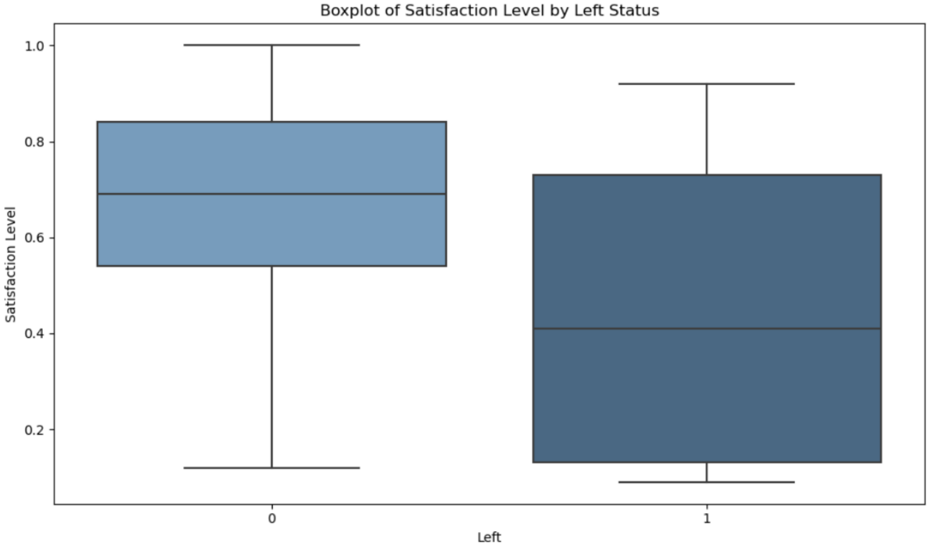
The bar graph titled "Retention Rate by Department" represents the retention rates of employees across different departments within an organization. The graph includes various departments such as IT, RandD, Accounting, HR, Management, Marketing, ProductMng, Sales, Support, Technical and The vertical axis shows the retention rate percentage. Each department has a corresponding bar that reflects its retention rate, with rates ranging approximately from 50% to 75%.

  - **Highest Retention Rates:** IT, RandD, and Management: These departments have the highest retention rates, indicating that employees in these departments are less likely to leave the company.
  - **Lowest Retention Rates:**Accounting and HR: These departments show lower retention rates, suggesting that employees in these areas are more likely to leave the company.
  - **Medium Retention Rates:** Marketing, Product Management (ProductMng), Sales, Support, and Technical: These departments have retention rates that are in between the highest and lowest.

Employees in these departments show average likelihoods of leaving the company compared to other departments.

- **Explore other relevant visualizations (e.g., histograms, box plots) to uncover insights related to retention.**

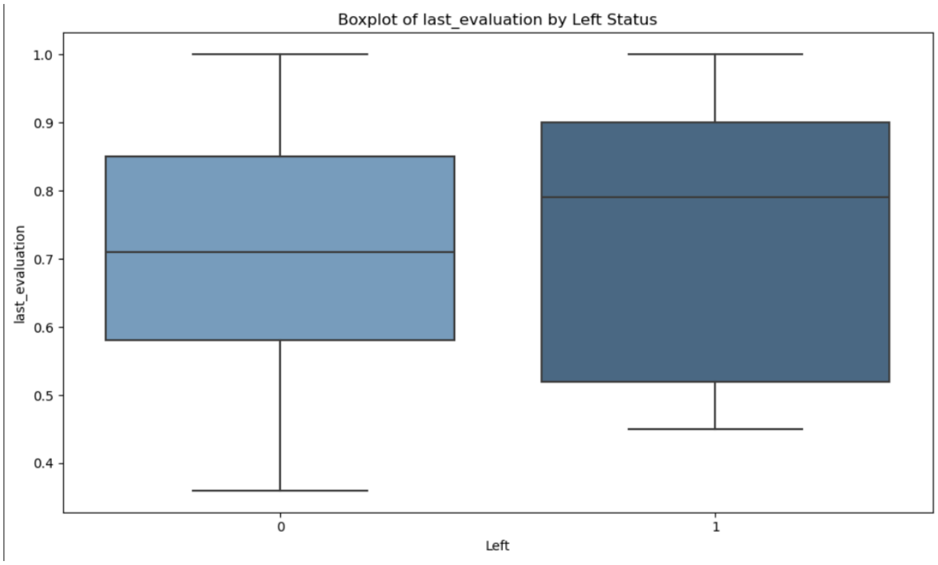
0.1 Retention based on vs.satisfaction levels:



The box plot compares the satisfaction levels of individuals based on their ‘Left Status’. The x-axis likely represents the ‘Left Status,’ with ‘0’ indicating those who stayed and ‘1’ indicating those who left. The y-axis represents the satisfaction level (usually ranging from 0 to 1).

- Higher Satisfaction for Stayers (Left Status 0): The boxplot shows that the median satisfaction level for individuals who stayed (Left Status 0) is higher. This suggests that those who remained in their roles or organization tend to be more satisfied.
- Lower Satisfaction for Leavers (Left Status 1): Individuals who left (Left Status 1) exhibit a lower median satisfaction level. This implies that those who departed were potentially less satisfied.

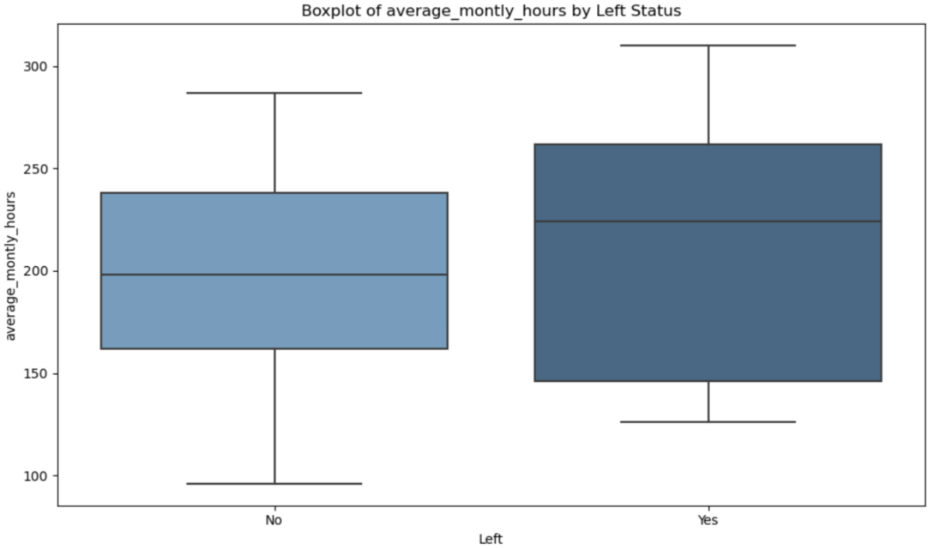
0.2 Retention based on last evaluation :



The boxplot of last evaluation by Left Status provides a visual comparison of the last evaluation scores between employees who have left and those who have not. Both groups have median lines near the 0.7 mark, suggesting similar evaluation scores on average. The group that has not left (Left Status 0) shows a wider range of scores, extending from lower to higher values. Also, The group that has left (Left Status 1) displays a more compact range of scores with no visible outliers.

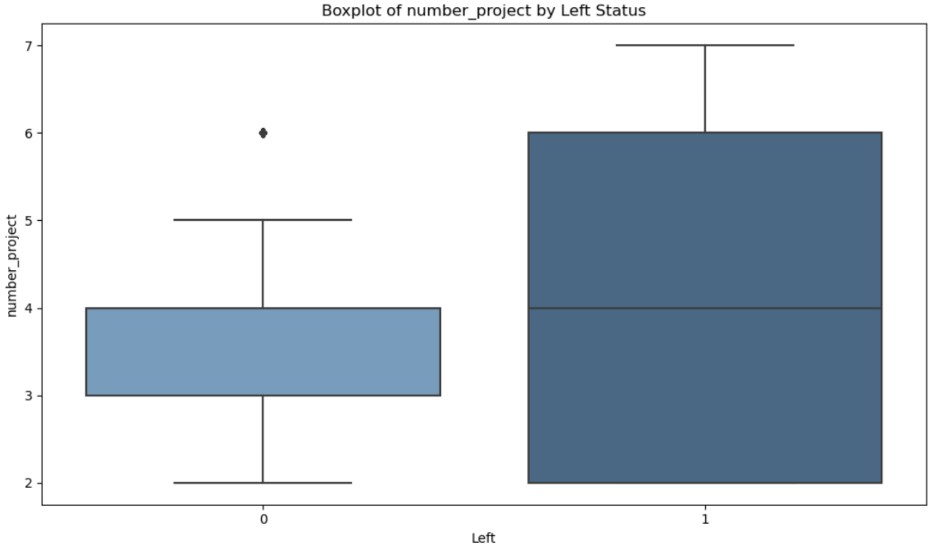
overall,the graph shows some overlap in evaluation scores between the two groups, it does not conclusively indicate that retention depends solely on the last evaluation. Other factors not shown in the graph may also influence an employee's decision to stay or leave.

0.3 Retention based on average montly hour :



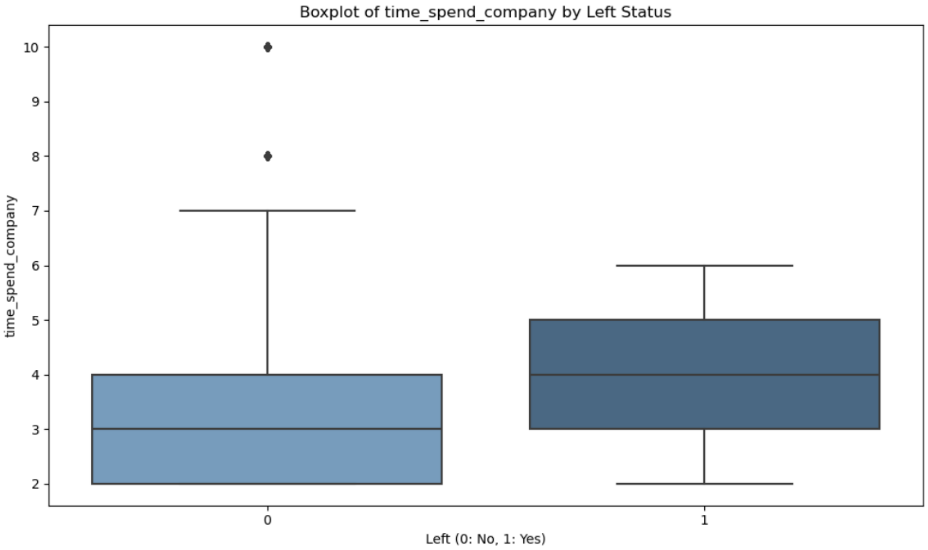
The boxplot graph provides insights into the relationship between average monthly hours worked and employee retention. Employees who have not left the company ('No') show a more symmetrical distribution of average monthly hours, suggesting a balanced workload. otherside, Lower Hours for Leavers: Those who left ('Yes') have a median closer to the lower end, indicating they worked fewer hours on average. Overall, There is greater variability in average monthly hours among employees who left, as shown by the wider interquartile range.

0.4 Retention based on number of project:



The "Boxplot of number project by Left Status" suggests a relationship between the number of projects employees handle and their retention. From the graph we can say that, Higher Project Count for Leavers mens Employees who left (Left Status 1) tend to have a higher median number of projects compared to those who stayed (Left Status 0). Overall, boxplot suggests that retention may be occure due to the higher number of projects .

0.5 Retention based on time spend comany:

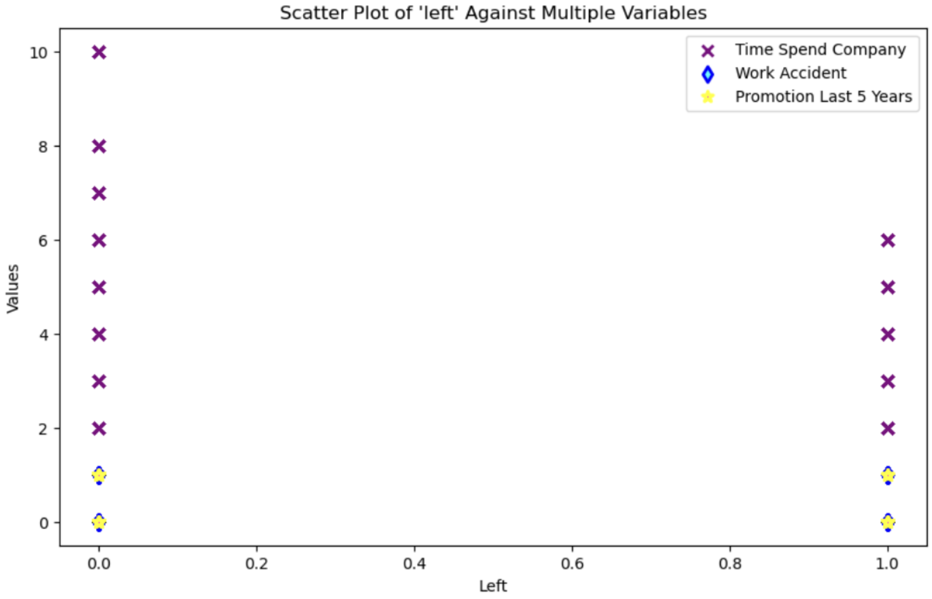


The boxplot compares the time spent at a company by employees who have stayed versus those who have left.

- Stayers (Left Status 0): The group that stayed with the company shows a smaller interquartile range (IQR), indicating a more consistent time spent at the company among these employees. Fewer outliers suggest that most stayers have similar employment durations. So, Employees who stayed (Left Status 0) show a more consistent time spent at the company.
- Leavers (Left Status 1): The group that left has a larger IQR, showing a wider variation in the time spent at the company. More outliers above the upper whisker indicate that some employees left after a significantly longer time than others. So, Employees who left (Left Status 1) display a wider range of time spent at the company.

Overall, boxplot suggests that retention may be influenced by the time employees spend at a company.

0.6 Retention based on Time Spent in Company, Work Accident, and Promotion in Last 5 Years:



The scatter plot shows the correlation between the 'left' vs. three other variables Time Spent in Company, Work Accident, and Promotion in Last 5 Years.

- Time Spent in Company : The crosses (x) are scattered across all 'left' values, indicating that the amount of time spent at the company can have a varying impact on employee retention.
- Work Accident : The diamond markers are grouped towards the lower end of 'left', suggesting that work accidents may have less influence on an employee's decision to leave.
- Promotion in Last 5 Years : The stars representing promotions are few but mostly found at lower 'left' values, suggesting that promotions could have a positive effect on retention but are not very common.

Feature Engineering:

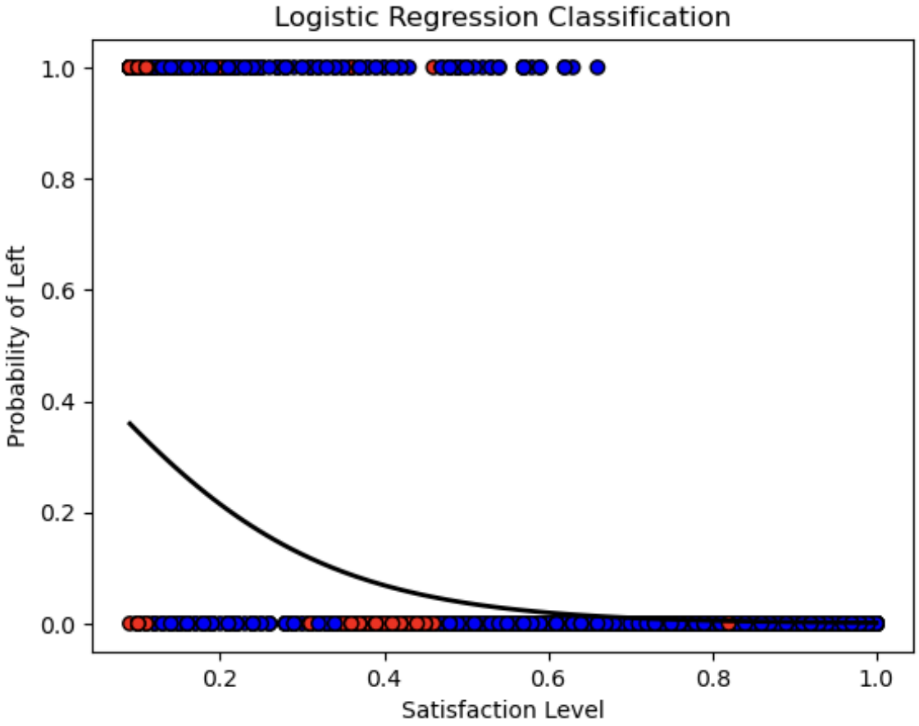
- **Based on the EDA, create new features that might be useful for the prediction model.**
- **Justify the choice of features selected for the model.**

  We can analyze the left and retention factors from the provided features, so there's no need to include any extra features.
- **Consider handling categorical variables, scaling numerical variables, and dealing with missing data appropriately.**

  Label encoding is used to transform categorical variables, such as salary and Department, into numerical values. Each unique category within a categorical variable is assigned a distinct integer label. Also, there is no missing values in the data.

Logistic Regression Model:

- **Split the dataset into training and testing sets.**
- **Build a logistic regression model using the variables identified in the EDA and feature engineering steps.**
- **Tune hyperparameters if necessary to improve model performance.**



The logistic regression graph shows relationship between a binary outcome (e.g., leaving or staying at a job) and a predictor variable (e.g., satisfaction level).

- Satisfaction Level is the predictor variable on the x-axis, ranging from 0 to 1.
- Probability of Leaving on y-axis
- Logistic Regression Line the black curve shows the predicted probabilities. It starts high when satisfaction is low and decreases as satisfaction increases.
- Data Points: The red dots represent actual observed data, showing where individuals with certain satisfaction levels ended up leaving or staying.
- The steepness of the curve indicates stronger relationship between satisfaction and leaving.

Model Evaluation:

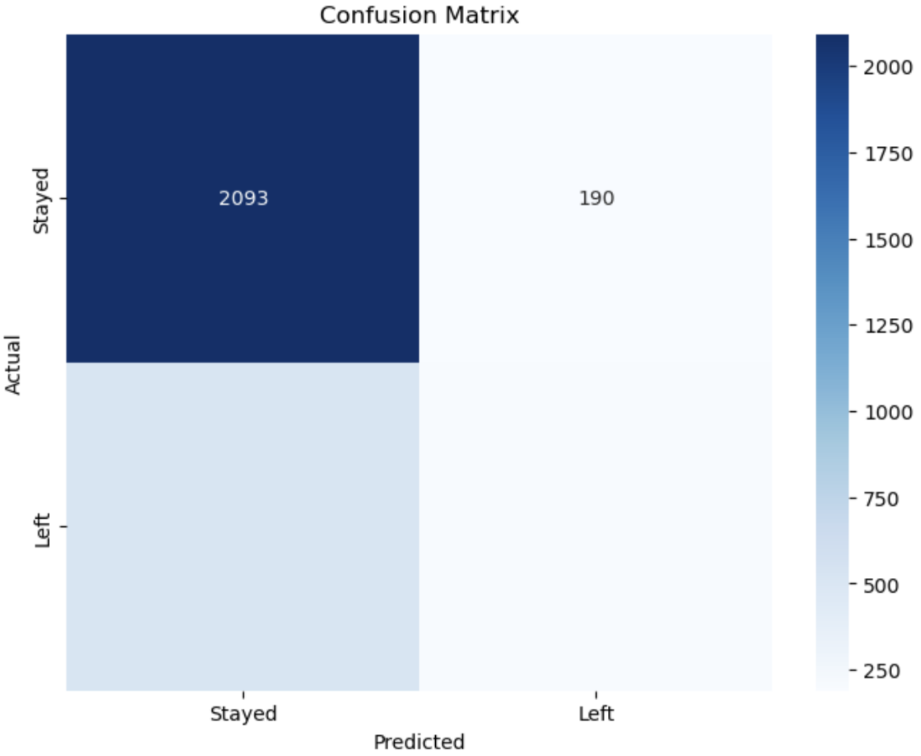
- **Measure the accuracy of the model on the test set.**
- **Compute and interpret other performance metrics such as precision, recall, F1 score, and ROC- AUC.**

  The performance metrics provided (Accuracy, Precision, Recall, F1 Score, and ROC AUC) are used to evaluate the effectiveness of your logistic regression model in predicting whether employees will leave the company

  - Accuracy : 0.7640, Accuracy is the proportion of correctly predicted instances (both true positives and true negatives) out of the total instances. It measures the overall correctness of the model. Here, accuracy of 76.40% means that the model correctly predicts whether an employee will leave or stay in about 76.40%of the cases. This is a decent level of accuracy, indicating that the model performs reasonably well.
  - A precision of 51.16% means that when the model predicts an employee will leave. This indicates that about half of the employees predicted to leave actually do leave. Precision is particularly important in contexts where the cost of false positives is high.
  - A recall of 27.75% means that the model correctly identifies the employees who actually leave. This is relatively low, indicating that the model misses a significant portion of employees who will leave (false negatives).
  - An F1 Score of 35.99% reflects a balance between Precision and Recall. Given that both Precision and Recall are not very high, the F1 Score is also relatively low, indicating room for improvement in the model's predictive power.
  - A ROC AUC of 0.7993 indicates that the model has a good ability to discriminate between employees who leave and those who stay. The value ranges from 0.5 (no discrimination) to 1 (perfect discrimination), so 0.7993 is considered quite good.
  - Summary: firstly, Accuracy is reasonably high, indicating overall good performance. Secondly, Precisionnindicates that about half of the positive predictions are correct, but this could be improved.

Thirdly, Recall is low, suggesting that the model misses a significant number of employees who actually leave. After that, F1 Score is relatively low due to the low Recall, indicating a need for better balance. Lastly, ROC AUC shows good discriminatory power, suggesting that the model is effective at distinguishing between employees who leave and those who stay.

- **Discuss the significance of each metric in the context of employee retention.**



The confusion matrix is a tool used to evaluate the performance of a classification model.

- True Positives (TP): The top-left cell (2093) represents the number of instances correctly predicted as ‘Stayed’.
- False Positives (FP): The top-right cell (190) represents the number of instances incorrectly predicted as ‘Stayed’ when they actually ‘Left’.

  - False Negatives (FN): The bottom-left cell (518) represents the number of instances incorrectly predicted as 'Left' when they actually 'Stayed'.
  - True Negatives (TN): The bottom-right cell (199) represents the number of instances correctly predicted as 'Left'.

Advanced Analysis:

- **Compare the logistic regression model with other classification algorithms (e.g., Decision Trees, Random Forest, Support Vector Machines).**
- **Perform cross-validation to ensure the robustness of the models.**

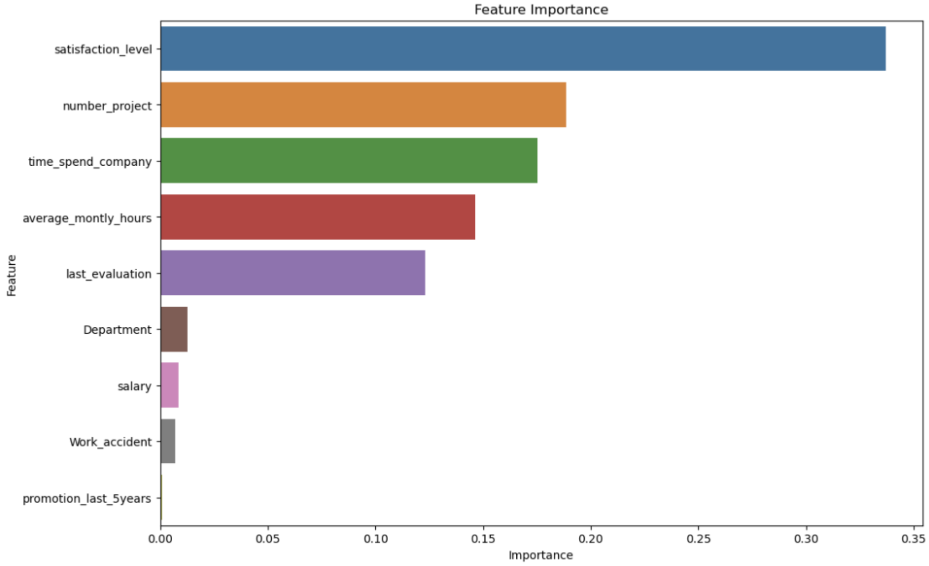
  Summary Based on the model accuracy results of different classification models and their cross-validation scores,

  - Logistic Regression: Accuracy: 0.7436 and Cross-Validation Accuracy: 0.7609 ± 0.0260. The logistic regression model performs moderately well, with a decent overall accuracy and some variability across different validation folds.
  - Decision Tree : Accuracy: 0.9748 and Cross-Validation Accuracy: 0.9815 ± 0.0135 The decision tree model shows very high accuracy, both in the initial run and cross-validation, indicating that it captures the patterns in the data well. However, it may be prone to overfitting.
  - Random Forest :Accuracy: 0.9921 and Cross-Validation Accuracy: 0.9931 ± 0.0084. The random forest model is the best performer with the highest accuracy and very low variability across validation folds, suggesting it generalizes well and captures complex patterns effectively.
  - Support Vector Machine (SVM) : Accuracy: 0.7839, and Cross-Validation Accuracy: 0.7845 ± 0.0035. So, The SVM model performs better than logistic regression but is outperformed by decision tree and random forest models. It has consistent performance across different validation folds.

As a result , Random Forest with an accuracy of 0.9921 and cross-validation accuracy of 0.9931 ± 0.0084, indicating exceptional performance and generalizability. While showing high accuracy, it might be less robust due to potential overfitting. Overall, Logistic Regression and SVM models perform reasonably well but are not as accurate or robust as the Random Forest model. So, the Random Forest model is recommended for predicting employee attrition due to its high accuracy and stability.

- **Conduct feature importance analysis to determine the most influential features for employee retention.**

- **Perform classification analysis to categorize employees into different risk levels of leaving the company (e.g., high risk, medium risk, low risk).**



The bar graph of "Feature Importance" shows the relative importance of different features in a dataset. The feature satisfaction level has the highest importance, suggesting it has a significant impact on the outcome being analyzed. Other Key Features 'number project', 'time spend company', and 'average monthly hours' are also important, indicating they are influential factors. Least Important Features 'Department' and 'salary' have low importance, which means they have minimal impact on the outcome.

Conclusion: The graph implies that employee satisfaction and workload-related features are crucial for the model's predictions, while department and salary are less critical.