

BIG DATA

MAY 2023

Big Data Programming Project:  
Attrition Analysis and Prediction using Machine Learning

Prepared by Ali ElGamal

Research Question:

How do different factors, such as age, distance from home, and frequency of business travel, impact employee attrition rates across different departments?

Different departments or job roles may have different attrition rates due to factors like work-life balance, job satisfaction, or career opportunities. By exploring how attrition rates vary across different departments or job roles, organizations can better understand which areas might be most at risk for high attrition rates and can take steps to address any issues.

this research question can be used to inform HR policies and practices. For example, if a particular factor like distance from home is found to be strongly associated with attrition rates, organizations could consider offering incentives like telecommuting or flexible work hours to help employees manage their work-life balance.

At the end of the project, the aim would be to identify which factors are most strongly associated with employee attrition rates and to develop recommendations for how organizations can improve employee retention. The results of this project could be used to inform HR policies and practices, to target retention efforts more effectively, and to ultimately improve employee satisfaction and productivity.

Data Scraping Technique:

For this project, I used web scraping to collect data from a website. I used Python's Beautiful Soup and Requests libraries to automate the process of accessing a table and extract all information and saving the data to a CSV file.

I extracted data from a table on a GitHub page containing employee attrition data. This is done using Python's Beautiful Soup and Requests libraries to automate the process of accessing a table and extract all information and saving the data to a CSV file

The dataset consists of 1470 observations and includes a dependent variable, "Attrition", and 31 independent variables. The independent variables are categorized into four main groups: Basic Info, Work Info, Satisfaction, Salary Related, and Time Related.

The Basic Info variables include Age, Gender, Education, Education Field, Martial Status, and Distance From Home.

Work Info variables include Department, Job Role, Job Level, Over Time, Business Travel, Performance Rating, Stock Option Level, and Job Involvement.

The Satisfaction variables include Work-Life Balance, Job Satisfaction, Relationship Satisfaction, and Enviroment Satisfaction.

Salary Related variables include Monthly Income, Monthly Rate, Daily Rate, Hourly Rate, and Percent Salary Hike.

The Time Related variables include Total Working Years, Traning Time Last Year, Years At Company, Years In Current Role, Years Since Last Promotion, Years With Current Manager, and Num Companies Worked.

Overall, this dataset provides a comprehensive view of various factors that may contribute to employee attrition, making it suitable for analysis to identify key drivers and potential solutions for retaining employees.

Data Processing type AND DIFFERENT BETWEEN EACH TYPE:

Parallel and sequential data processing are two different approaches to processing data. In sequential processing, data is processed one step at a time, with each step being completed before moving on to the next. This approach is often used in applications where the processing steps are dependent on the output of the previous step.

On the other hand, parallel processing involves dividing the data into smaller chunks and processing them simultaneously using multiple processing units or processors. This approach is often used in applications where large volumes of data need to be processed quickly.

There are several advantages to using parallel processing over sequential processing. For example, parallel processing can significantly reduce the time required to process large datasets, as it allows multiple processing units to work on different parts of the dataset simultaneously. This can result in faster processing times and improved performance.

Another advantage of parallel processing is that it can help to reduce the risk of data loss or corruption. When processing large datasets, the risk of errors or data loss can increase, particularly if the processing is being done on a single processing unit. By using parallel processing, the workload can be distributed across multiple processing units, reducing the risk of data loss or corruption.

However, parallel processing also has some disadvantages. For example, it can be more complex and difficult to implement than sequential processing, as it requires specialized hardware and software. Additionally, parallel processing can be more expensive than sequential processing, as it requires additional hardware and infrastructure.

In my project , I used sequential data processing for the data analysis. This approach was suitable for the dataset size and complexity we were working with. However, for larger datasets or more complex analysis, parallel processing may be a more appropriate approach.

Performance analysis

One essential component of performance analysis is measuring the time it takes for a code block to execute. This is commonly referred to as "time execution" and can be accomplished using various tools and techniques.

n this report, we will focus on measuring the time execution of code blocks using the Python library pandas in a Jupyter Notebook. We will use the time module to capture the start and end times of the code block and calculate the time difference. By doing so, we will be able to determine the performance of our code and identify potential areas for optimization.

Additionally, we will also introduce memory execution tests, which allow us to measure the memory consumption of a code block. This can help us identify memory leaks and optimize our code for better memory usage

1. Data Scraping Performance analysis

A screenshot of a computer

Description automatically generated with low confidence

The output above shows the results of the time and memory execution tests for the data scraping code.

The "Elapsed time" line indicates that it took 11.23 seconds for the entire code block to run, from the point where the timer started to the point where it ended. This gives an idea of the efficiency and speed of the code.

The "Current memory usage" and "Peak memory usage" lines indicate the amount of memory used by the code during its execution. In this case, the current memory usage at the end of the code block was 63.870542 MB and the peak memory usage during its execution was 64.025529 MB. This gives an idea of the memory requirements of the code and can help identify any potential memory leaks or inefficiencies.

2)Correlation Heatmap Performance Analysis



The output shows the results of performance analysis for a correlation heatmap using the seaborn library. The elapsed time for generating the heatmap was 0.55 seconds, which is significantly faster than the data scraping task. The current memory usage and peak memory usage are also much lower for this task, indicating that it requires less memory to run.

The performance of a task can depend on various factors such as the complexity of the code, the amount of data being processed, the algorithms and libraries used, and the hardware and software environment.

3)Multi-job role bar charts



For this task, the elapsed time was 6.81 seconds, the current memory usage was 2.838441MB, and the peak memory usage was 3.716805MB. This implies that creating bar charts for each job role in the dataset took a moderate amount of time and memory usage.4) machine learning models accuracy comparisons:

A picture containing text, font, screenshot

Description automatically generated

This code computes and prints the accuracy scores of four different machine learning models, which are Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting.

The output of the code displays the accuracy scores for each model, which indicate how well each model performed at classifying the data. A higher accuracy score indicates that the model is better at correctly classifying the data.

The Logistic Regression model achieved an accuracy score of 0.8469, which means it correctly predicted the target variable for 84.69% of the samples. The Decision Tree model achieved an accuracy score of 0.9374, which is higher than the Logistic Regression model, indicating that it performed better in predicting the target variable. The Random Forest model achieved the highest accuracy score of 0.9599, indicating that it had the best performance among the models tested. Finally, the Gradient Boosting model achieved an accuracy score of 0.9224, indicating that it performed better than Logistic Regression, but not as well as the Decision Tree and Random Forest models.

Overall, the Random Forest model had the best performance among the four models tested, achieving an accuracy score of 0.9599.

Data Visualization:

As we can see from this figure that there is no data missing. Where If there are no missing values in the DataFrame, then the output of this code will be a series of zeros, one for each column in the DataFrame. If there are missing values, the output will show the number of missing values in each column.  
A screenshot of a computer

Description automatically generated

Data Preprocessing:

I removed some low-correlated columns and kept some high-correlated ones.The figure below shows the table with the selected features and their label encoding. I also cleaned the data by dropping unnecessary columns, checking for missing values and duplicates. I reduced the data dimensionality and numerosity by selecting highly correlated features.

Feature selection is a crucial step of pre-processing because it affects the model performance. Some features are less relevant and some are more relevant for the context. The model will perform better with better features.

A picture containing text, screenshot, font

Description automatically generated

The figure above shows the dropped low-correlated columns.

A picture containing text, screenshot, rectangle, diagram

Description automatically generated

As mentioned earlier the data set comprises of 35 columns and 1470 rows. The "Attrition" column in the dataset has two possible values: "Yes" and "No," where "Yes" represents employee attrition and "No" represents the opposite. Out of the total observations, there are 237 "Yes" responses and 1233 "No" responses, indicating that the dataset is imbalanced, with a significantly higher number of observations belonging to the "No" class. The attrition rate can be calculated as 237/1470, which is approximately 16.1% as the graph shows

A picture containing screenshot, text, diagram, rectangle

Description automatically generated

Correlation analysis involves measuring the strength of the relationship between two sets of variables, which can be independent or dependent. The strength of this relationship is represented by the correlation coefficient, which can be positive or negative. In machine learning algorithms, accuracy may not be the best metric to evaluate performance, so correlation plots can be useful for revealing relationships between variables. Redundant features can be avoided by using features that are not highly correlated when developing predictive models. Heatmap plots provide a visual representation of the pairwise correlations between variables, and the code in this context identifies and prints pairs of variables with correlation coefficients above 0.7. The output shows that job level and total working years, years at the company and years with current manager, and years at the company and years in current role have strong positive correlations. The heatmap also reveals a strong positive correlation between monthly income and job level. In summary, the heatmap provides insight into the correlations between variables related to employee job satisfaction.

The above heatmap suggests that many columns have a weak correlation with each other. It is preferable to use features that are not highly correlated when developing a predictive model to avoid redundant features. Techniques such as Principal Component Analysis (PCA) can be used to reduce the feature space if there are many correlated features.

Looking at the Heatmap, we can see that there is a strong positive correlation between job level and total working years, with a correlation coefficient of 0.78. This suggests that employees who have been with the company for a longer time tend to have higher job levels.

Another strong positive correlation is between years at the company and years with current manager, with a correlation coefficient of 0.77. This indicates that employees who have been at the company for a longer time tend to stay with their current manager.

There is also a strong positive correlation between years at the company and years in current role, with a correlation coefficient of 0.76. This suggests that employees who have been at the company for a longer time tend to stay in their current role.

Additionally, there is a strong positive correlation between monthly income and job level, with a correlation coefficient of 0.95. This indicates that employees with higher job levels tend to have higher monthly incomes.

Overall, this heatmap shows the strength and direction of correlations between different variables related to employee job satisfaction

Exploratory Data Analysis:

I performed exploratory data analysis on the IBM HR dataset to identify the columns that had high correlation and to determine which columns I used to build a machine learning model comparison. I created charts that shows the attrition rates against different factors.

A picture containing text, screenshot, rectangle, plot

Description automatically generated

As you can see from the bar chart above, The gender distribution plot by attrition count It shows that the males have more attrition cases than females .

A picture containing text, screenshot, diagram, plot

Description automatically generated

We can see in the above figure that the "Research & Development" department has the highest employees, and the lowest number of employees who left are from the "HR" department.

A picture containing text, screenshot, diagram, plot

Description automatically generated

The bar chart shows the attrition against age group. Age is grouped from 20-3-,30-40,40-50 and 50-60. The 30-40 age groups showed the highest amount of attrition which is 85. While similarly the age group 20-30 had very similar attrition count which is 84. Also the lowest attrition count was at the age group 50-60. This shows that with increasing age, the attrition rate began to fall as people sought stability in their jobs.

I prepared the data for the ML model by encoding the categorical features (using LabelEncoding or OneHotEncoding) to make them numerical. I also scaled the features to the same range by using the StandardScaler from scikit library.

Machine learning model:

The output shows the performance metrics of four different models: Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting. Each model has an accuracy, precision, recall, F1-score, and confusion matrix associated with it.

The performance metrics of each model are then shown. The accuracy represents the overall performance of the model, while precision and recall show how well the model classifies positive and negative examples. F1-score is the harmonic mean of precision and recall, which gives a balance between the two. Finally, the confusion matrix shows the number of true positives, true negatives, false positives, and false negatives.

Conclusion

The machine learning models applied to the IBM HR dataset showed the following results:

The accuracy score of each model is as follows:

• Logistic Regression Accuracy: 0.8469387755102041

• Decision Tree Accuracy: 0.9374149659863945

• Random Forest Accuracy: 0.9598639455782313

• Gradient Boosting Accuracy: 0.9224489795918367

The accuracy score represents the percentage of correctly classified instances out of all instances in the dataset. A higher accuracy score indicates better performance of the model.

In this case, the Random Forest model has the highest accuracy score (95.99%), indicating that it performs the best among the four models on the given dataset. Moreover, we think that machine learning models can be used by businesses in the future to forecast employee attrition. This has a direct practical benefit because an HR department can plan the optimal workforce for each department. Calculate the number of employees needed for each department so that the recruitment process can be done smoothly and effectively. These case studies supported the research question by showing that ensemble methods with effective feature selection are successful in forecasting employee attrition, as shown by visualizations and accuracies by different models, and thus managers should focus on the top needs of employees by motivating entry-level employees, enhancing job satisfaction, and relationship satisfaction.