Predicting the Attrition Rate of a Company

Manasi Todankar
Student, School of Information Studies
Syracuse University
Syracuse, 13210
mptodank@syr.edu
https://www.youtube.com/watch?v=ds3zYkO7Q1c

ABSTRACT

Attrition refers to when an employee is leaving the organization in contrast to turnover where a company is successful in replacing the leaving employee. Employee Attrition is one of the most significant problems faced by organizations around the world. The primary goal of this research is to create a model that can predict whether an employee would quit the organization. This research paper also tries to shed some insight on the many elements influencing worker attrition and their potential remedies. Implementing this approach will aid management in employee evaluation and decision-making by identifying valuable individuals who will depart the organization. Using this program, hidden reasons for employee attrition may be identified, and management can take preventive measures for each employee's attrition individually.

Keywords: Machine Learning, Employee Attrition, Classification, Prediction, Random Forest Tree, Decision Tree, Logistic Regression, Gaussian Naïve Bayes, KNN, SVM.

1 INTRODUCTION

Skilled individuals move from job to job, taking with them the customer knowledge and technical experience that the organization requires. Their pay is rising, as are their perks, benefits, and bonuses. This study is being conducted to identify the most influential factors driving attrition and the prediction of attrition.

1.1 Why do employees leave?

To enable a company to establish and implement a successful retention plan, senior and line management must first identify the factors that cause top performers to depart and seek alternative jobs. Employees experience one or more of the these when they think about their job: undervalued, underwhelmed, underpaid, or overworked. "If employees don't get along with their bosses, don't like them, or don't respect them, they will leave a firm despite a high pay or wonderful perks," write Marcus Buckingham and 5Curt Coffman. A competent boss, regardless of remuneration, will instill loyalty." Good employees depart because they feel they will be treated better by another organization.

As well as several other elements like as work environment, compensation, welfare, working hours, job advancement, and personal/family are the reasons for the intention of leaving.

An organization must devote a significant amount of time and money in training each employee in accordance with the firm's

needs. When an employee quits the organization, the corporation not only loses its important people, but it also loses the money it invested to recruit, select, and train those individuals for their particular professions. On the other hand, the business must continue to spend in the recruitment, training, and development of new employees in order to fill vacant roles. Because of these factors, every corporation strives to reduce attrition and retain employees through improving corporate rules and work conditions.

Machine learning techniques are employed in this study to solve the problem of employee attrition for two reasons. To begin, machine learning techniques have not been applied to alleviate staff retention concerns in recent years. Second, machine learning approaches surpass the attrition problem in employees.

2 METHODOLOGIES

Very first step, data was acquired from free online web source and predictive analysis techniques have been used on this data.

Dataset: EDA-Analyzing the Attrition Rate of a Company

Link:

https://www.kaggle.com/code/muhammedsal98/eda-analyzing-the-attrition-rate-of-a-company/notebook

2.1 Data description

This dataset contains 29 columns and 4410 rows. 'Attrition' column is our target feature. It is a Binary Classification type as it contains 'Yes' or 'No'. The other columns will act as the factors determining what led the employee to resign or retire.

Table 1: Attribute Description

EmployeeID	Identification number of the employee
Age	Age of the employee
Attrition	Whether the employee left in the previous year or not
BusinessTravel	Frequency of business travel in the last year
Department	Department in the company
DistanceFromHome	Distance from home in km
Education	Education Level: 1 'Below College', 2 'College', 3 'Bachelor', 4 'Master', 5 'Doctor'
Education Field	Field of education
EmployeeCount	Employee count
Gender	Gender of employee
JobLevel	Job level in the company on a scale of 1 to 5
JobRole	Name of job role in the company
MaritalStatus	Married, Single or other
MonthlyIncome	Monthly income in rupees per month
NumCompaniesWorked	Total number of companies the employee has worked for
Over18	Whether the employee is above 18 years of age or not
PercentSalaryHike	Percent salary hike for last year
StandardHours	Standard hours of work for the employee
StockOptionLevel	Stock option level of the employee
TotalWorkingHours	Total number of years the employee has worked so far
TrainingTimesLastYear	Number of times training was conducted for this employee last year
YearsAtCompany	Total number of years spent at the company by the employee
YearsSinceLastPromotion	Number of years since last promotion
YearsWithCurrManager	Number of years under current manager
EnvironmentSatisfaction	Work Environment Satisfaction Level 1 'Low' 2 'Medium' 3 'High' 4 'Very High'
JobSatisfaction	Job Satisfaction Level: 1 'Low', 2 'Medium', 3 'High', 4 'Very High'
WorkLifeBalance	Work life balance level 1 'Bad' 2 'Good' 3 'Better' 4 'Best'
JobInvolvement	Job Involvement Level 1 'Low' 2 'Medium' 3 'High' 4 'Very High'
PerformaceRating	Performance rating for last year 1 'Low' 2 'Good' 3 'Excellent' 4 'Outstanding'

2.2 Data Exploration

Exploring data is the most important part of understanding any data. Statistical techniques and visualization are used to understand the analytical action of the data. By analyzing the data, we understand the important aspects of the data. For this research paper, data exploration in perform in the form of exploring it in a statistical way. Also, visualizations such as bar plots, scatterplots and density plots are used. Using these visualizations have helped to establish relationships between two variable and give an overview on how they might be affection the overall prediction of attrition in the company.

2.2.1 Descriptive Statistics

Descriptive statistics are necessary to understand the numerical data. It helps us in identifying any abnormality in the data briefly rather than checking the data individually. Here, the data is accurate and there are no entry errors.

Table 2: Descriptive Statistics

	EmployeeID	Age	DistanceFromHome	Education	EmployeeCount	JobLevel	MonthlyIncome	NumCompaniesWorked
count	4410.000000	4410.000000	4410.000000	4410.000000	4410.0	4410.000000	4410.000000	4410.000000
mean	2205.500000	36.923810	9.192517	2.912925	1.0	2.063946	65029.312925	2.694830
std	1273.201673	9.133301	8.105026	1.023933	0.0	1.106689	47068.888559	2.493497
min	1.000000	18.000000	1.000000	1.000000	1.0	1.000000	10090.000000	0.000000
25%	1103.250000	30.000000	2.000000	2.000000	1.0	1.000000	29110.000000	1.000000
50%	2205.500000	36.000000	7.000000	3.000000	1.0	2.000000	49190.000000	2.000000
75%	3307.750000	43.000000	14.000000	4.000000	1.0	3.000000	83800.000000	4.000000
max	4410.000000	60.000000	29.000000	5.000000	1.0	5.000000	199990.000000	9.000000

PercentSalaryHike	StandardHours	...	TotalWorkingYears	TrainingTimesLastYear	YearsAtCompany	YearsSinceLastPromotion	YearsWithCurrManager
4410.000000	4410.0	...	4410.000000	4410.000000	4410.000000	4410.000000	4410.000000
15.209524	8.0	...	11.279936	2.799320	7.008163	2.187755	4.123129
3.659108	0.0	...	7.774275	1.288978	6.125135	3.221699	3.567327
11.000000	8.0	...	0.000000	0.000000	0.000000	0.000000	0.000000
12.000000	8.0	...	6.000000	2.000000	3.000000	0.000000	2.000000
14.000000	8.0	...	10.000000	3.000000	5.000000	1.000000	3.000000
18.000000	8.0	...	15.000000	3.000000	9.000000	3.000000	7.000000
25.000000	8.0	...	40.000000	6.000000	40.000000	15.000000	17.000000

EnvironmentSatisfaction	JobSatisfaction	WorkLifeBalance	JobInvolvement	PerformanceRating
4410.000000	4410.000000	4410.000000	4410.000000	4410.000000
2.723603	2.728246	2.761436	2.729932	3.153741
1.089654	1.098753	0.703195	0.711400	0.360742
1.000000	1.000000	1.000000	1.000000	3.000000
2.000000	2.000000	2.000000	2.000000	3.000000
3.000000	3.000000	3.000000	3.000000	3.000000
4.000000	4.000000	3.000000	3.000000	3.000000
4.000000	4.000000	4.000000	4.000000	4.000000

2.2.2 Data Cleaning

The dataset may contain null values, outliers, and redundant columns. Before we do any data exploration, we need to clean the data. First, I have checked foe null values in the columns of the dataset. The null values are present in the numerical columns of the dataset. So, I chose to replace them with the mean of the respective columns they are present in. In descriptive statistics it is clear that the data is normal and does not have to be treated for any outliers. All the data entries in the dataset make sense. Next I have changes the values of certain columns that contain special characters just for the ease of further data preprocessing.

Also, it is necessary to get rid of any columns that might not be helpful in building a machine learning model. Thus, I have dropped the columns named 'Over18', 'EmployeeCount', StandardHours', and 'EmployeeID' because either they have only one unique value or they don't have importance to make a difference further on.

2.2.3 Visualization

To get the complete overview we can compare the count of 'Yes' to 'No', ie. The number of employees that left the company last year versus the the emplyoyees that decided to stay.

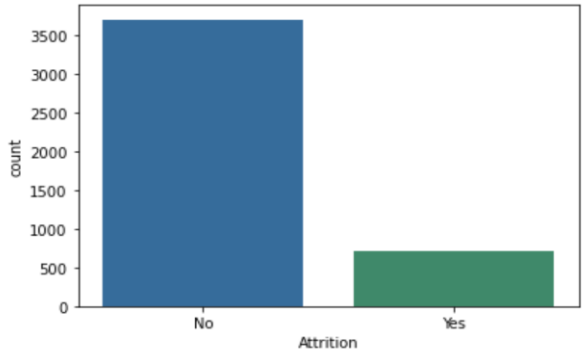


Exhibit 1: Attrition comparison

We can observe that compared to the number of employees who decided to stay, the number of employees leaving are quite low.

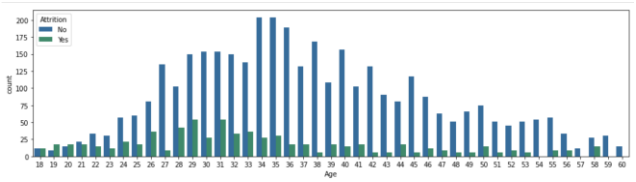


Exhibit 2: Age wise distribution

Here we can see that roughly 24 years to 36 years is the range where employees end up leaving the company.

Below are some more visualizations in attempt find patterns or relationships within the data.

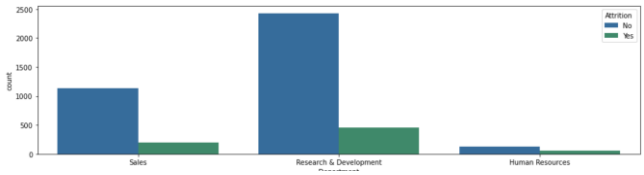


Exhibit 3: Department wise distribution

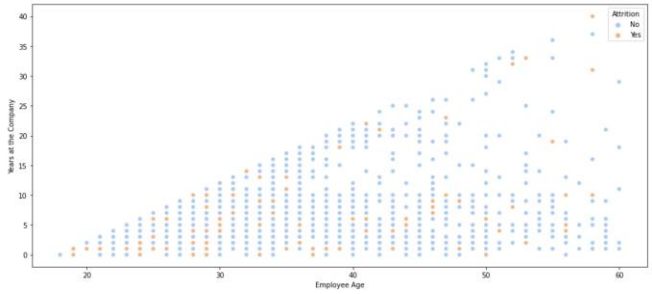


Exhibit 4: Relation between age and the number of years spent at the company

This scatterplot is an attempt to link the age of an employee to the number of year he/she spent at the company. It gives us a chance to understand if the employees are leaving that approached their working term or leaving much before that time.

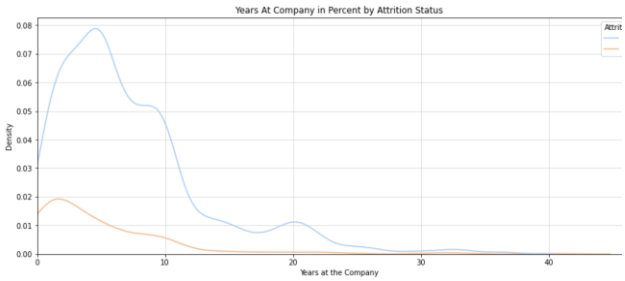


Exhibit 5: Years at the Company by Attrition

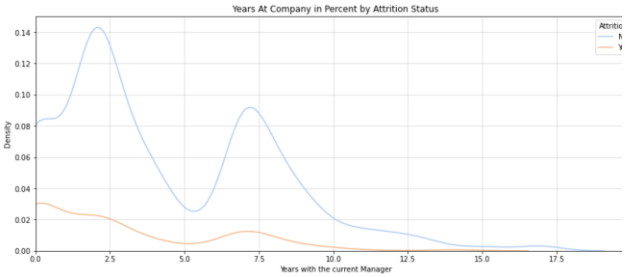


Exhibit 6: Years with current Manager

The above two density plots give us an overview of years spent at the company and years spet with the current managers. These visualization may lead us to explore the working environment within the company.

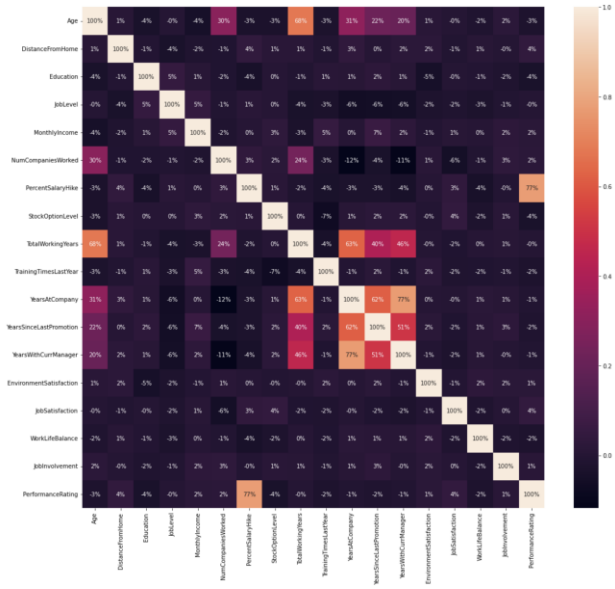


Exhibit 7: Correlation between different features

4 DATA MODELLING

I have divided the data in 25% test data and the rest train data.

3.1 Modelling

The modeling process involves selecting models based on various machine learning approaches to be employed in experimentation. Various predictive models based on artificial neural networks, DT, Bayesian technique, logistic regression, SVM, and so on can be used in prediction. Our objective is to find the best classifier for our situation. Each classifier may be tracked on the feature set for this, and the classifier with the best classification results can be utilized for prediction.

After comparing the accuracy scores of various models, I decided to use Random Forest Tree algorithm and Decision Tree Algorithm.

3.1.1 Random Forest Tree

Random Forest Tree is an ensemble learning approach for categorizing and backsliding the dataset. This approach works while outputting the mode of the classes (categorizing) or backsliding of the particular tree by developing a large number of Decision Trees.

3.1.2 Decision Tree

A Decision Tree, breakdown the decisions which visually and clearly signify decisions and decision making. A Decision Tree describes data in which the resulting classification tree can act as an input for decision making.

3.2 Training and Testing

Using cross-validation, the dataset is divided into training and testing datasets, with training data used to train the model and testing data used to test the model.

Decision Tree (DT), and Random Forest were the machine learning techniques employed for the prediction model (RF). Scikit-learn was used to train the classifiers.

3.3 Model Evaluation

Metrics for the evaluation of a classification model

Accuracy:

The accuracy returns the proportion of correct predictions.

Precision:

The precision returns the proportion of true positives among all the values predicted as positive.

Recall:

The recall returns the proportion of positive values correctly predicted.

Specificity:

The specificity returns the proportion of negative values correctly predicted.

F1-score:

The f1-score is the harmonic mean of precision and recall. It is often used to compare classifiers.

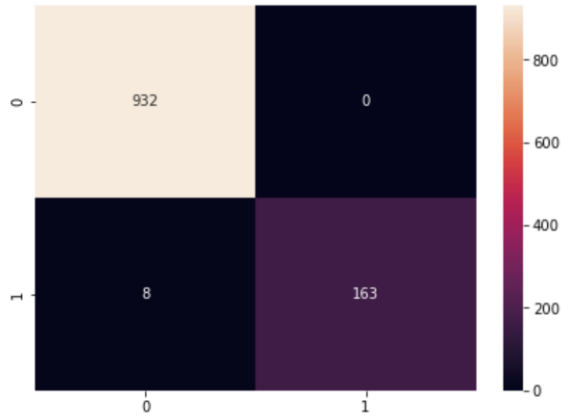
The harmonic mean gives more weight to the lower value, so a high F1-score means that both precision and recall are high.

Support: number of observations for each class.

Macro average: the arithmetic average of a metric between the two classes.

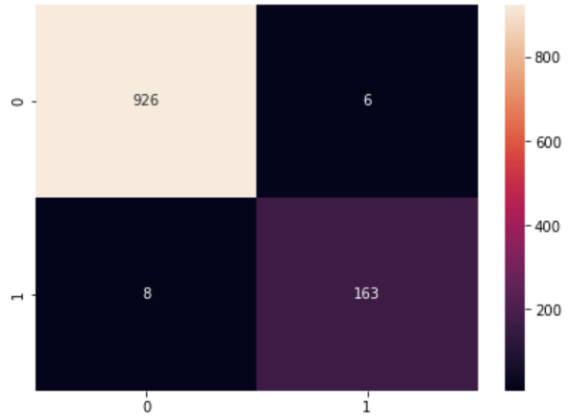
Weighted average: the weighted average is calculated by dividing sum(metric of interest x weight) by sum(weights).

3.2.1 Random Forest Tree Model



	precision	recall	f1-score	support
0	0.99	1.00	1.00	932
1	1.00	0.96	0.98	171
accuracy			0.99	1103
macro avg	1.00	0.98	0.99	1103
weighted avg	0.99	0.99	0.99	1103

3.2.2 Decision Tree Model



	precision	recall	f1-score	support
0	0.99	0.99	0.99	932
1	0.95	0.95	0.95	171
accuracy			0.98	1103
macro avg	0.97	0.97	0.97	1103
weighted avg	0.98	0.98	0.98	1103

4 RESULTS

When the classifiers are evaluated using the confusion matrix, it is discovered that the Random Forest achieves even greater accuracy than the Decision Tree, surpassing

all of the classifiers. One reason RF outperforms Decision Tree is that Decision Tree utilizes the whole sample in each step, selecting decision boundaries at random rather than selecting the best one. It is clear from the findings, as Random Forest has a 99% accuracy rating. The accuracy of Decision Tree and Random Forest is substantially higher, and it appears that these classifiers may be used to forecast if an employee would quit the organization.

5 CONCLUSION

In the case of employee attrition, an estimate may be made as to whether the person will quit the company or not. Using this technique, the company may identify employees who are most likely to leave and then provide them limited incentives. There may also be certain situations of false positives, in which the HR believes an employee will leave the firm soon, but the individual does not. These errors may be costly and inconvenient for both employees and human resources, but they are a better bargain for relational growth. On the other side, there might be a false negative if a human resource does not promote or raise personnel and they quit the firm.

The employee attrition prediction problem is concerned with people's decisions. Various machine learning algorithms were used to the human resource dataset in this paper. Based on the findings of this study, it is possible to infer that Random Forest algorithm outperforms.

REFERENCES

Jain, P.K., Jain, M. & Pamula, R. Explaining and predicting employees' attrition: a machine learning approach. SN Appl. Sci. 2, 757 (2020).

Zhao, Y., Hryniewicki, M.K., Cheng, F., Fu, B., Zhu, X. (2019). Employee Turnover Prediction with Machine Learning: A Reliable Approach. In: Arai, K., Kapoor, S., Bhatia, R. (eds) Intelligent Systems and Applications. IntelliSys 2018. Advances in Intelligent Systems and Computing, vol 869. Springer, Cham. https://doi.org/10.1007/978-3-030-01057-7_56

Attrition, Turnover & Retention: What's the difference? Plus How to Calculate Attrition? (2019, April 29). [Video]. YouTube. https://www.youtube.com/watch?v=l27ngeU

Dr. S. Rabiyathul Basariya, & Dr. Ramyarrzgarahmed. (2019). A STUDY ON ATTRITION – TURNOVER INTENTIONS OF EMPLOYEES. International Journal of Civil Engineering and Technology (IJCIET).

G.R., K.A., K.L., K.N.S.G., & M.S.K. (2022). Employee Attrition Prediction using Machine Learning. *IRJMETS*.

Laurenti, G., PhD. (2021, December 16). *Confusion Matrix and Classification Report - The Startup*. Medium. https://medium.com/swlh/confusion-matrix-and-classification-report-88105288d48f