Employee Attrition Enigma

Pooja B. Baba
Computer Science
Georgia State University
Atlanta, Georgia
pbaba1@student.gsu.edu

Saloni R. Sawal
Data Science in Analytics
Georgia State University
Atlanta, Georgia
ssawal1@student.gsu.edu

***Abstract*— Employee attrition implies not only the loss of an employee but also the loss of a customer from an organization. A higher rate indicates a failure of organizational efficiency in terms of retaining skilled employees. The purpose of this project is to perform Data Analysis on the 'Employee Attrition' dataset and compute the probability of attrition using various algorithms. The results thus obtained will be used to understand what changes the company should make to their workplace, to get most of their employees to stay. To achieve this, we have implemented machine learning models. Machine-learning approaches are the most effective instruments for achieving this goal. In this study, we analyzed three Machine Learning classifiers, such as Logistic Regression, Random Forest Classifier, and Decision Tree Classifier, and selected the model which is the best to ameliorate employee attrition by calculating accuracy, precision, recall, and f1-score.**

***Keywords*—employee attrition, classification, machine learning**

I. Introduction

Employee attrition is unavoidable in any business. However, if the situation is not handled properly, the departure of key employees can lead to a decrease in productivity. The organization may have to hire new employees and train them on the tool being used, which will take time. Most organizations want to know which of their employees are on the verge of leaving.

In HR practice, the term Employee Attrition is interchangeable concerning industry and its causes. Attrition means reducing an employee through retirement, resignation, or death. In most of the research, it has been found that work-related is the primary cause of a higher employee attrition rate. High attrition results in a loss in the company's cost spent on recruitment and training. The impact of employee attrition leaves a long-term negative impression on the organization's goodwill. In simple words, it can be said that employee attrition is caused due to nonfulfillment of an employee's perception or expectation towards the employer or failure of the employer's commitment to employees' satisfaction.

Cause of Employee Attrition: The below image displays the most common reasons noted so far for employees leaving the firm.

This paper discusses the application of the Logistic Regression, Random Forest, and Decision Tree algorithm as a method of predicting employee attrition. This is done by using data from Kaggle and
treating the problem as a classification task. The conclusion is reached by comparing the performance of the algorithm.

II. Previous Work

Employee Attrition Analysis Using Predictive Techniques[1] is a springer conference paper published in 2017, written by Devesh Kumar Srivastava and Priyanka Nair. This paper represents predictive analytics techniques and implements the framework to predict the employee attrition

Early Prediction of Employee Attrition using Data Mining Techniques[2] is an IEEE paper published in 2018, written by Sandeep Yadav, Aman Jain, and Deepti Singh. This paper considers employee behavior and implements various data mining and machine learning algorithms to determine employee attrition beforehand. The data referred to by this paper was the CompData[3] surveys conducted.

III. Data Analysis

Since one of our goals is to predict the chance of Attrition, let's take a look at how the different variables correlate with it.

1) *Inference*-The below visualization shows the attrition count for different job levels**.** Yes, means the number of employees stays in the company and No means number of employees left the company.

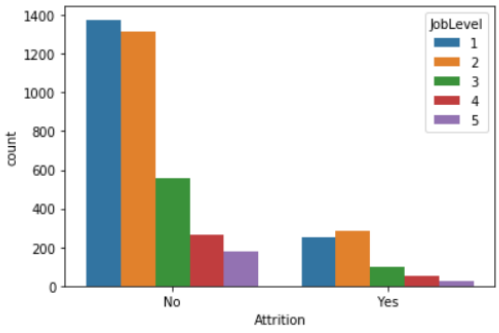

Figure 1 Attrition count for different job levels

As you can see in the below chart, around 16% of the employees have left the company.

Attrition	Percentage
Yes	0.84
No	0.16

2) Job satisfaction level of employees: -

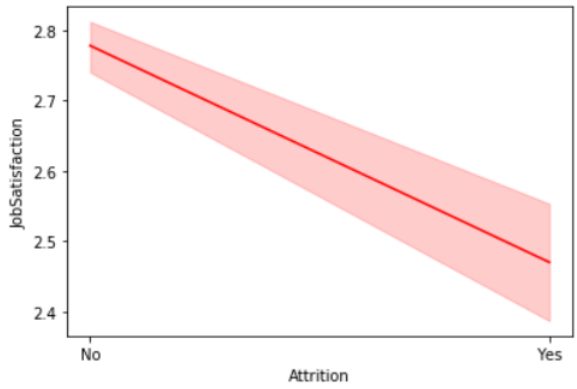

Figure 2 Job satisfaction level of employees

Inference- It is observed that the chances of an employee with less job satisfaction at the workplace leaving the company is high.

3)Promotion effects on employees: -

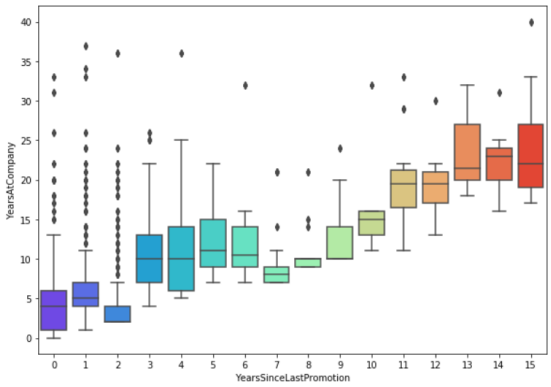

Figure 3 Promotion effects on attrition

Inference - It is observed that for employees who spend more time in the company their chances of getting promoted are decreased.

4) Attrition Rate by Marital Status: -

Relation	Percentage
Divorced	10.09
Married	12.48
Single	25.53

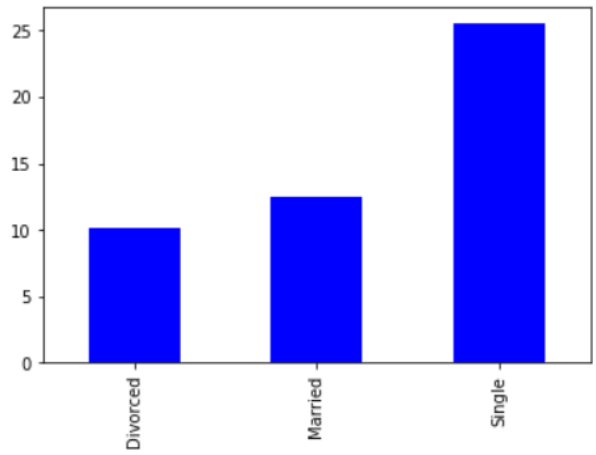

Figure 4 Attrition based on marital status

Inference - As shown in the above figure, single people tend to be attracted towards employee attrition concept more as compared to divorced and married people.

5) Monthly Income Comparison: -

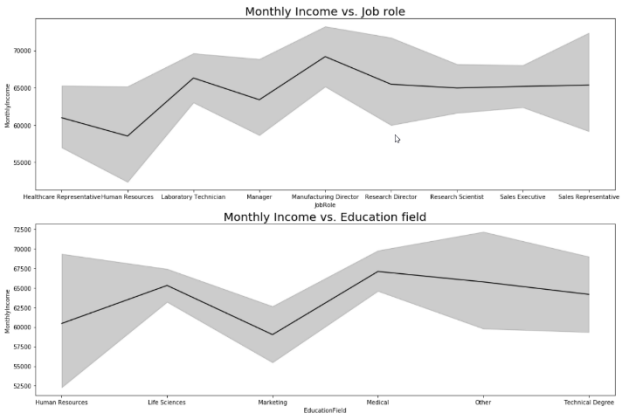

Figure 5 Monthly income comparison

Inference-From the above plots, it is evident that an employee with a job role as a manufacturing director or education field as medical gets the highest salary whereas, an employee with a job role in HR or education field as marketing gets the lowest salary.

IV. PROPOSED FRAMEWORK FOR EMPLOYEE ATTRITION PREDICTION

1. *Logistic Regression*

Logistic regression is a classification model that fits the values of the logistic function. Logistic Regression is a statistical model to model the relationship between input variables and output variables It is useful when the dependent variable is categorical. The general form of the model is

$$P(Y|X, W) = 1\ 1+e\ -(w0+\sum wi\ xi\ )$$

Logistic regression is often used with regularization techniques to prevent overfitting.

2. *Random Forest Model*

In Random Forests, we choose a random selection of features for constructing the best split and use an ensemble to import it. Ensemble learning methods are made up of a set of classifiers e.g., a decision tree, and their predictions are aggregated to identify the most popular result.

It's a kind of ensemble technique that combines a bunch of weak models to create a powerful model. The random forest generates several tresses. Voting for that class should be used to categorize each tree. This is a categorization. The forest chooses the classification with the most votes.

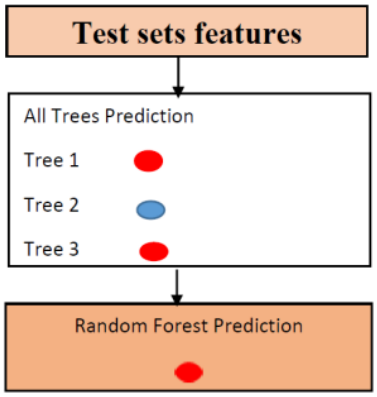

Figure 6 . Prediction Process taken by random forest

3) *Decision Tree Classifier*

The Decision Tree algorithm identifies various ways of splitting data into branch-like segments. It partitions data into subsets based on categories of input variables. Its primary purpose is to create a training model that may be used to anticipate employee attrition decisions using data sets from previous studies. It attempts to tackle the problem by using nodes or node hierarchies.

The root node represents the complete sample, separated into leaf nodes that display the attribute divided into leaf nodes representing the class Labels.

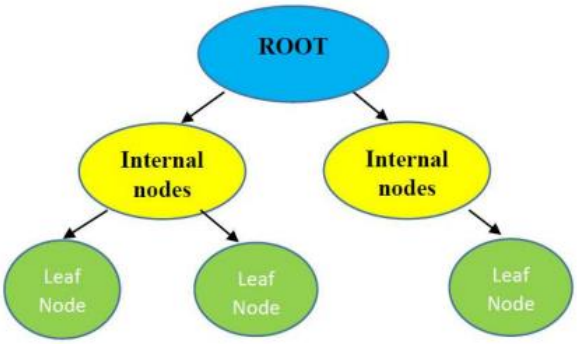

Figure 7 Representation of decision tree

Evaluation

1. *Accuracy*

Accuracy is an evaluation metric that measures the total number of predictions that a model gets right.

Accuracy = Correct Predictions / Total Predictions

2. *Precision Score*

Precision refers to the number of positive class predictions that belong to the positive class. It can be seen as a measure of quality.

Precision = TruePositive / (TruePositive + FalsePositive)

3. *Recall*

Recall refers to the percentage of total relevant results correctly classified by the algorithm. It can be seen as a measure of quantity.

Recall = TruePositive / (TruePositive + FalseNegative)

V. IMPLEMENTATION

The dataset used for the implementation of this project was from Kaggle[4]. The dataset had the following files – general_data.csv, employee_survey_data.csv, in_time.csv, out_time.csv, and manager_survey_data.csv. *general_data.csv* file consisted of the basic information of the employee such as age, salary, marital status, job level, gender, etc. There are a total of 24 columns in this file. i*n_time.csv* & *out_time.csv* these files consist of the timestamp at which the employee enters or leaves the firm for the day. m*anager_survey_data.csv* file consists of the survey data carried out by the employees about their managers. c*mployee_survey_data.csv* consists of survey data carried out by the employers.

This csv data was loaded in the data frame using python's pandas library. To train any machine learning models on this loaded data, data pre-processing was required. Data Cleaning was performed on the 2 columns – NumberOfCompaniesWorked and TotalWorkingYears. The null values in these columns were represented with the median values of the corresponding columns.

To understand the data better, various data visualizations were used. Heatmap, column chart, stacked column chart, box-plot, pair plot, line plot, joint plot. were used. These plots also helped to check for any anomalies.

For training purposes, Logistic Regression, Random Forest Classifier, and Decision Tree classifier models were used. Using python's sklearn library the model classes were chosen and were trained on the data. The data was split into training and testing data using sklearn's *train_test_split* function. For model selection following evaluation, metrics were used – precision, f1-score, recall.

VI. Result

Classification Report: -

The performance metrics like Classification Accuracy, Precision, and Recall are considered in this paper to evaluate the performance of the feature selection methods in the prediction of employee attrition in the industry using different Machine Learning classifiers.
Following are the metrics for the implemented models –
1) *Random Forest Classifier*

	Precision	Recall	F1-score	Support
0	0.86	1.00	0.92	1493
1	0.96	0.09	0.17	271
Accuracy			0.86	1764
Macro avg	0.91	0.55	0.55	1764
Weighted avg	0.87	0.86	0.81	1764

2) *Decision Tree Classifier*

	Precision	Recall	F1-score	Support
0	0.87	0.97	0.92	1493
1	0.54	0.17	0.26	271
Accuracy			0.85	1764
Macro avg	0.70	0.57	0.59	1764
Weighted avg	0.82	0.85	0.82	1764

3)*Logistic Regression*

	Precision	Recall	F1-score	Support
0	0.86	0.98	0.92	1493
1	0.58	0.15	0.24	271
Accuracy			0.85	1764
Macro avg	0.72	0.57	0.58	1764
Weighted avg	0.82	0.85	0.81	1764

VII. Conclusion

Employee attrition prediction has become a vital issue in today's organizations. Employee attrition is a significant problem for businesses, especially when trained, technical, and critical staff leaves for better opportunities elsewhere. This leads to a financial loss as a qualified employee must be replaced. This paper presented why prediction is essential. It further outlined various classification algorithms based on supervised learning to solve the prediction problem. We have performed all the pre-processing steps required. After pre-processing, we used this pre-processed dataset to implement our machine learning models. In this project, we have Implemented three machine-learning algorithms to predict employee attrition in an organization and used three metrics, Precision score, Recall, and F1-Score, to evaluate the performance of each of our models. Among the models implemented, all models give us a similar accuracy of 0.85, we will use Random Forest Classifier to predict the attrition for a sample employee**.**

VIII. Acknowledgment

We thank Dr. Yanqing Zhang for his guidance, encouragement, and co-operation throughout the completion of this paper.

IX. References

1. Employee Attrition Analysis Using Predictive Techniques- https://link.springer.com/chapter/10.1007/978-3-319-63673-3_35
2. Early Prediction of Employee Attrition using Data Mining Techniques - https://ieeexplore.ieee.org/abstract/document/8692137/authors#authors
3. CompData Survey Results - https://www.servicenow.com/lpayr/employee-experience-survey.html?campid=83737&cid=p:hr:dg:nb:rmkt:exa:goog_hrsm_restructure:ams:all&s_kwcid=AL!11692!3!586315159575!p!!g!!employee%20survey&ds_c=GOOG_AMS_All_EN_DEMANDGEN_HRSM_RLSA_NonBrand_PHR_Other-res&cmcid=71700000091578400&ds_ag=Employee+Survey_PHR&cmpid=58700007713043216&ds_kids=p70017734770&gclid=Cj0KCQjw37iTBhCWARIsACBt1IybD8GFA1ij96ZzdnykGAWi3XeGOOlqd8UT6UTk6yXjOWT0XlG5Yo4aAg98EALw_wcB&gclsrc=aw.ds
4. Kaggle dataset - https://www.kaggle.com/vjchoudhary7/hr-analytics-case-study