Employee Attrition Prediction in HR Analytics

Aafrith U.M.

EG/2020/3795
Department of Computer Engineering
University of Ruhuna,Faculty of Engineering
Galle,Sri Lanka
aafrith_um.ruh.ac.lk

Ahamed M.F.R.

EG/2020/3807
Department of Computer Engineering
University of Ruhuna,Faculty of Engineering
Galle,Sri Lanka
ahamed_mfr.ruh.ac.lk

***Abstract*:**

Employee attrition is the important thing in most companies. The departure of employees from a company makes difficulties for businesses in the present day. It results in disruptions in work processes, reduced efficiency and higher expenses linked to the recruitment and training of staff members. To solve these challenges effectively, companies need some methods to reduce employee turnover. This document gives a system for predicting employee attrition in HR analytics using machine learning algorithms such as Naive Bayes and Logistic Regression algorithms. Its purpose is to identify employees who're at risk of leaving the organization.

Keywords—Attrition, Regression, Machine Learning, Preprocessing, Histogram, Piechart, Data analysis, Employee, Prediction, Box plot, Encoding

I. INTRODUCTION

Employee attrition, which is most of the employees are leaving companies creates a main problem, for businesses in the day. It can cause so many problems, in the workflow, decreased productivity, and increased costs associated with hiring and training new employees. The financial impact of turnover can be quite significant often amounting to a portion of an employees salary.

To solve the issue of employees leaving the company we implementing a machine learning based employee attrition prediction system using Naive Bayes, Logistic Regression algorithms. This system identifies the reasons such, as job satisfaction, work life balance and compensation to predict the probability of an employee leaving. Through the identification of possible leaving employees, Companies may begin measures for maintaining talent and reducing employee turnover.

This project is very important for the companies who are facing this problem. These are some of the importance that is mentioned here: (1) Saving the costs: Companies have to find new employees and train them will cost some financial amount. For that this predicting and preventing employee attrition is very important to save the financial resources. (2)Productivity will improve: Retaining skilled and experienced employees are very much needed for a company to increase productivity and the performances of the company. (3)Maintaining Talents : Companies may use some strategies and methods to keep on valued talent by identifying employees who are at the risk of leaving. (4) Enhanced Motivation among Staff: A stable employee with low attrition rates creates a friendly and positive work atmosphere that is helpful to both the employees and the company.

We hope to achieve the main things from this project are: (1)Have to identify the main factors which are causing employee attrition.(2)Have to develop a machine learning model that can accurately predict employee attrition.(3)To give companies information that will help in reducing attrition and helping the companies keep valued staff.

We have chosen two main algorithms from machine learning to do this project effectively. They are:(1)Naive Bayes : Naive Bayes is a classification algorithm. It is very simple and good efficiency algorithm when it is used with the datasets. It assumes that features are conditionally independent which makes it a good fit, for our problem. In our case we have features related to employee attributes and job satisfaction.(2)Logistic Regression : It is also a classification algorithm. It is known to be understandable and having the ability to handle datas and numerical elements. We can determine the relative importance of multiple variables in predicting employee attrition through the use of logistic regression.

There are mainly some advantages and disadvantages of these two algorithms. Advantages :(1) Both algorithms are efficient, can handle large datasets and making them suitable for many applications.(2) These algorithms are widely used in many fields to get the results very effectively. Disadvantages :(1)Overfitting and underfitting is a problem when we use this two algorithms.(2)Selecting the features is very important because it can effect the performance of both algorithms.

Our aim is to create a good and efficient employee attrition prediction system by carefully identify these possible disadvantages and using the advantages of the selected algorithms.

.

II. METHODOLOGY

A. *Description of Dataset*

The dataset used in this project was obtained through Kaggle, a popular data science and machine learning platform [1]. An employee is represented by each row. A variety of features, including monthly income, performance rating, work-life balance, environment satisfaction, and job satisfaction, are included in each column. There are about 1470 data points available for 35 features total. Our main goal is to turn employee attrition into a task of classification to predict it. This indicates that our goal is to predict if a worker will quit (attrition = 1) or remain (attrition = 0). Our methodology is based on this dataset, and it allows us to use algorithms to find patterns and learn more about employee turnover in a company's environment.

B. *Pre-processing*

During the preprocessing stage of our project, we carefully reduced the size and improved the quality of our dataset so that it could be used for analysis. We carefully removed features like StandardHours, EmployeeCount, Over18, EmployeeNumber, and StockOptionLevel after observing how unimportant some features were. After carefully checking the data for duplicates, we were certain that there were no duplicate data.

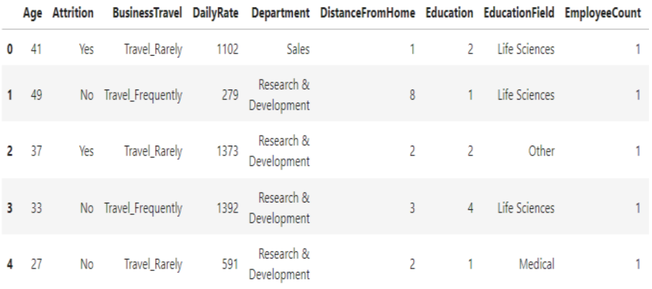
In addition, a thorough examination for null values produced a clean dataset free of any missing data. We used LabelEncoder to convert categorical variables which included features like Attrition, Business Travel, Department, EducationField, Gender, JobRole, MaritalStatus, and OverTime into numerical categories in order to allow an additional analysis. This makes the categorical data suitable with our models of analysis while maintaining its integrity. We used Minmax scaling to normalize our data according with standard methods, ensuring accurate and reliable input for following analysis. Our dataset is now prepared and cleaned as a result of these good preprocessing methods, creating the foundation for considerable and accurate findings into employee attrition prediction.

C. *Algorithms*

The classification algorithms that are used in the project for predicting employee attrition was logistic regression and Naive Bayes. (1)By using Logistic regression: Using a set of data, the model was trained to identify patterns and connections between different features. An evaluation of the model's testing and training accuracy showed a significantly higher test accuracy of 88.86% and a train accuracy of 86.93%. Evaluation criteria such as F1-score, precision, and recall highlighted the model's benefits and weaknesses. (2)By using Naïve Bayes : To predict employee attrition, the another algorithm that is used the Naive Bayes classifier, which is the Gaussian Naive Bayes implementation from the scikit-learn toolkit. With a train accuracy of 77.86% and a test accuracy of 80.71%, the model showed its capacity to identify similarities in the data. The factors that measured the model's performance included accuracy, recall, and F1-score.

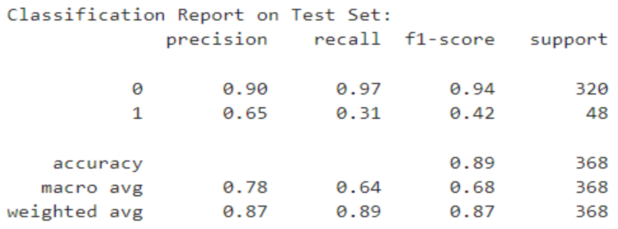
D. *Implementation*

Here is the implementation of our project using these 2 algorithm. A Logistic Regression model was developed using scikit-learn library, achieving 86.93% train and 88.86% test accuracy. The model was evaluated using precision, recall, and F1-score metrics. The Naive Bayes approach, trained on the dataset, achieved 77.86% train and 80.71% test accuracy. Both models demonstrated their ability to identify patterns in the dataset, providing valuable insights into potential employee attrition scenarios. The comparative analysis of these algorithms enhances their applicability in organizational context, enabling informed decision-making regarding employee retention strategies.



	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount
0	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1
1	49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	1
2	37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	1
3	33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	1
4	27	No	Travel_Rarely	591	Research & Development	2	1	Medical	1

Figure 1 : Sample of our data set description

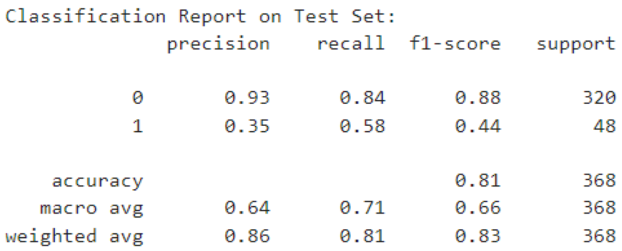


```
Classification Report on Test Set:
              precision    recall  f1-score   support

           0       0.90      0.97      0.94       320
           1       0.65      0.31      0.42        48

    accuracy                           0.89       368
   macro avg       0.78      0.64      0.68       368
weighted avg       0.87      0.89      0.87       368
```

Figure 2 : Logistic regression classification report from confusion matrix



```
Classification Report on Test Set:
              precision    recall  f1-score   support

           0       0.93      0.84      0.88       320
           1       0.35      0.58      0.44        48

    accuracy                           0.81       368
   macro avg       0.64      0.71      0.66       368
weighted avg       0.86      0.81      0.83       368
```

Figure 3 : Naive Bayes classification report from confusion matrix

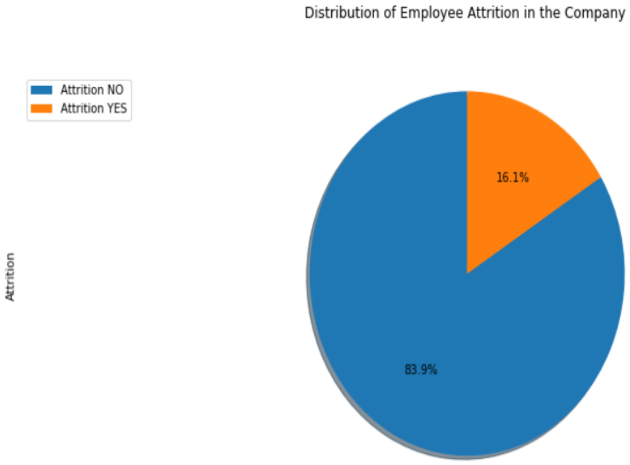


Figure 4 : Employee Attrition rates analysis of the company

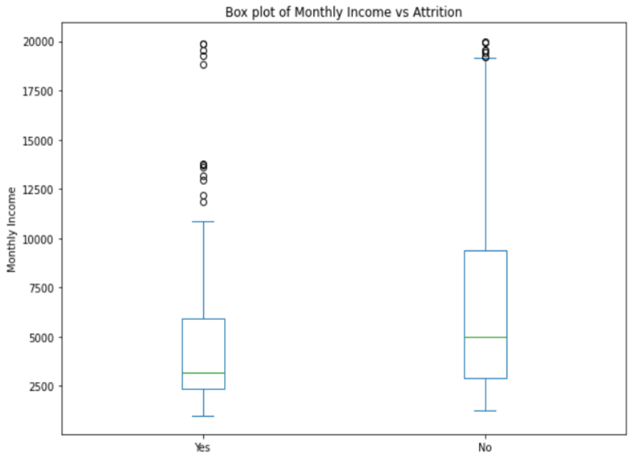


Figure 5 : Box plot analysis of the Monthly income vs attrition

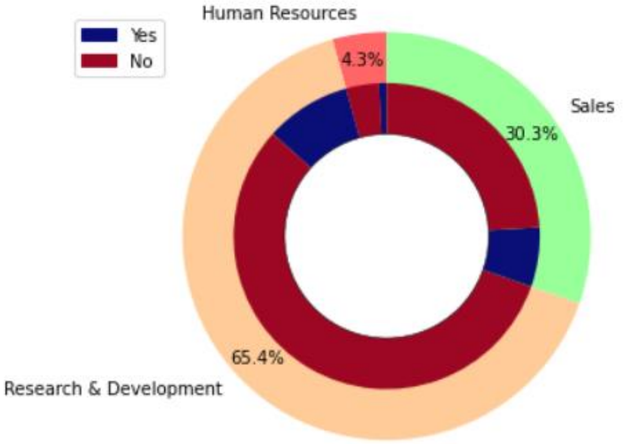


Figure 6 : department feature analysis

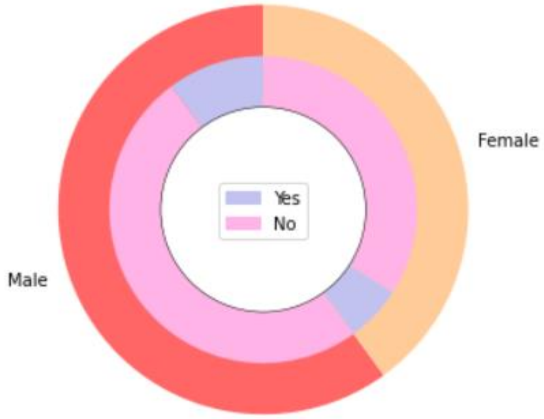


Figure 7 : Gender Analysis

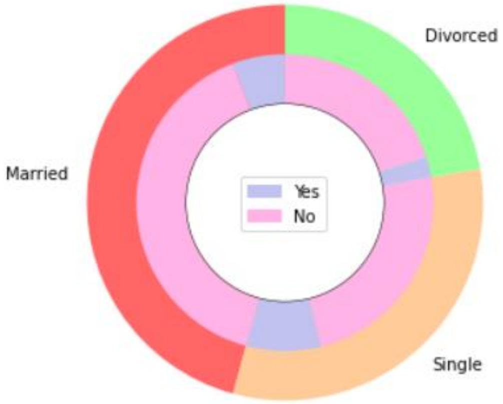


Figure 8 : Marital status analysis

III. Results

The machine learning models, Logistic Regression and Naive Bayes Regression, were evaluated based on key performance metrics to assess their effectiveness in predicting employee attrition rates.

Model	Logistic Regression	Naïve Bayes Regression
Accuracy	0.89	0.81
Precision	0.87	0.86
Recall	0.89	0.81
F1-score	0.87	0.83

The results indicate that the Logistic Regression model outperforms the Naive Bayes Regression model across all considered metrics. The higher accuracy, precision, recall, and F1-score of the Logistic Regression model suggest its superior ability to predict employee attrition.

IV. Discussion and conclusion

A. Discussion:

In comparison of the two models Logistic Regression always exhibits greater accuracy and precision than Naive Bayes regression. This implies that Logistic Regression is better able to determine and categorize workers who are likely to leave, giving critical information for strategic retention initiatives.

Indeed, although both models exhibit proficiency, Logistic Regression's capability to capture subtle patterns in the data makes it a better performer. Nevertheless, one should be aware of the limitations and potential biasness that can come about in any machine learning based model.

B. Ethical Aspects:

Attention to ethical treatment of employee private information, including raise data, is a priority for this project; privacy and confidentiality measures are built into the process. This is consistent with ethical values and judicial regulations that seek to maintain high standards of privacy protection in predicting employee turnover.

We highlight transparency and informed consent by giving information to participants about why we use their data for attrition prediction. This strengthens trust and aligns with ethical principles that emphasize informed consent.

To be fair and not biased, our project also acts by minimizing unfairness in the training data since it is designed so as never discriminate towards any group. This commitment to fairness in model design also ensures responsible and equitable use of machine learning for employee attrition prediction.

C. Conclusion:

The findings suggest that machine learning, especially Logistic Regression, works well when used in the prediction of employee attrition rates. These models provide a valuable tool for organizations that seek data-based insights into their work force dynamics.

The importance of Logistic Regression is significant to an organization since it has high accuracy and higher precision that helps the organizations take proactive steps in retaining valuable talent. The results should add to the emerging field of HR analytics and workforce management.

References

[1] "Deloitte," [Online]. Available: https://www.deloitte.com/global/en/our-thinking/insights/topics/talent/content/human-capital-trends.html. [Accessed 21 January 2024].

[2] "scikit-learn," scikit-learn, [Online]. Available: https://scikit-learn.org/stable/index.html. [Accessed 21 January 2024].

[3] V. Ferreira, "ResearchGate," December 2018. [Online]. Available: https://www.researchgate.net/publication/331074766_Motivacao_para_o_Desporto_em_Atletas_Jovens_Motivation_for_Sports_in_Young_Athletes_pp_135-143. [Accessed 21 January 2024].

[4] PAVANSUBHASH, "Kaggle," Kaggle, 7 year ago. [Online]. Available: https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset. [Accessed 21 January 2024].

[5] A. P. Gulati, "Analytics Vidhya," 17 November 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/11/employee-attrition-prediction-a-comprehensive-guide/. [Accessed 21 January 2024].

[6] "Scikit learn," [Online]. Available: https://scikit-learn.org/stable/modules/naive_bayes.html. [Accessed 21 January 2024].