HR analytics: Employee attrition analysis using logistic regression

I Setiawan*, S Suprihanto, A C Nugraha and J Hutahaean
Department of Computer Engineering and Informatics, Politeknik Negeri Bandung, Bandung, Indonesia

*irwan@jtk.polban.ac.id

Abstract. Employee attrition can become a serious issue because of the impacts on the organization’s competitive advantage. It can become costly for an organization. The cost of employee attrition would be the cost related to the human resources life cycle, lost knowledge, employee morale, and organizational culture. This study aimed to analyze employee attrition using logistic regression. The result obtained can be used by the management to understand what modifications they should perform to the workplace to get most of their workers to stay. The data for the study were around four thousand employees, covering 261 days (one year working days) during 2015 — the data period between January and December. We use R for data integration, exploratory data analysis, data preparation, logistic regression, model evaluation, and visualization. The study has five steps: (1) data collection and business understanding, (2) data pre-processing, (3) exploratory data analysis, (4) model selection and training, and (5) test and evaluation of the model. The result of the study found eleven variables as key driving factors for employee attrition. It also showed that the model has 75 percent accuracy with 73 percent sensitivity and 75 percent specificity.

1. Introduction
The purpose of Human Resource Management (HRM) is covering worker performance and commitment, analyzing workers collaboration models, investigating employee churn and turnover, and creating employee lifetime estimation. Human resource analytics is a developing application area of analytics of HRM objectives [1]. On the other hand, the reputation of HR Analytics restricted in a high-quality scientific evidence-based study [2]. The employee is one of the most critical assets in today’s knowledge-driven industry. Employee attrition can become a serious issue because of the impacts on the organization’s competitive advantage. Employee attrition can become costly for an organization. The cost of employee attrition would be the cost related to the human resources life cycle, lost knowledge, employee morale, and organizational culture.

In 2018, Mitchell and Steven, using employees’ questionnaire answers regarding their manager, perform research to measure people’s management skills. They investigate the intensity to which people management skills correlate to employee outcomes, with the most comprehensive focus on employee attrition. The principal conclusion is that people management has a strong negative relationship with employee attrition [3]. Frye, Boomhower, and Smith present a model for predicting employee attrition. They used a collection of statistically significant factors that relate to an employee's choice to resign. They applied Principal Component Analysis and K-Nearest Neighbors and Random Forest. They found that Logistic
Regression predicts employee leaves with the highest accuracy of their testing methods, reaching a higher than seventy-four percent success rate [4].

In 2016, Raja and Kumar evaluated the circumstances that influence employee attrition in the IT division. They used a questionnaire for data gathering with a sample of 300 workers in the IT division. The instruments they used are simple percentage method, chi-square method, and correlation coefficient method. They found that there is no correlation between factors that retain the employee in the IT division and the factors that influence the employee work environment [5]. Bindra, Sehgal, and Jain, analyzed Employee Attrition data to predict the employee attrition based on five classes. They use the Association Rule Algorithm ‘Apriori’ and Decision Tree Algorithm ‘C5.0’. The conclusion is that the performance of the algorithm is increased when they use C5.0 with the association rule algorithm as compared to using C5.0 independently [6]. In 2016, Mishra and Lama proposed a model to improve the effectiveness and performance of the Human Resource system that will optimize business outcomes by providing a framework for Human Resource decision-making related decisions comprising data mining and predictive analytics [7].

Regression methods have become an integral part of any data analysis, which concerned with explaining the correlation between a response variable and one or more analytical variables. The logistic regression model is the most frequently used regression model for the analysis of these data [8]. Logistic Regression has been used to identify the factors that determine job satisfaction [9], work-life balance [10], work environment and production loss [11], stress, and dissatisfaction on employees [12], and incentive effects on employee engagement [13].

This paper aims to analyze employee attrition using logistic regression. The result thus obtained can be used by the management to understand what modifications they should perform to the workplace to get most of their workers to stay.

2. Methods

2.1. Experiment tool
We use R studio for data integration, exploratory data analysis, data preparation, logistic regression, model evaluation, and visualization.

2.2. Attrition analysis steps
The analysis process is divided into five steps (see Figure 1), as follows:

- Data Collection and Business Understanding: collecting and analyzing the data to grasp better what should be the main goals of the study.
- Data Pre-processing: pre-process the data to suit them with the analysis method. The pre-processing may involve cleaning up the data, transforming the data, or creating new variables that may bring useful information for the analysis steps.
- Exploratory Data Analysis (EDA): this step create textual and visual summaries of the dataset that highlight some characteristics of the data.
- Model Selection and Training
- Test and Evaluate the Model: evaluate the performance of the proposed models

Figure 1. Attrition analysis steps.
3. Results and discussion

3.1. Data collection and business understanding
The data are covering 261 days (one year working days) during 2015. The data period between January and December. We observed 4,410 employees. The firm is divided into three departments and classified by job roles. The core department at the firm is Research & Development, comprising 65 percent of employees. The percentages of the employee education level of College and below College is 31 percent. The education field of the firm dominates by Life Sciences and Medical with 41% and 32 %, respectively. The regular working hours are 7-8 hours per day. The attrition rate for this data is 15.70 percent.

The problem definition is how to model the probability of employee attrition using logistic regression. The result thus obtained can be used by the management to understand what modifications they should perform to their workplace to get most of their workers to stay.

The data are divided into five datasets: emp_general, emp_manager_survey, emp_survey, emp_intime, and emp_outtime. The emp_general is about general employee data; emp_manager_survey consist data about survey result of job involvement and performance rating conducting by the manager; emp_survey consist data about environment satisfaction, job satisfaction, and work-life balance survey results — attendance data stored in two separate datasets, which are emp_intime and emp_outtime. Figure shows the number of objects and variables for each data.

![Figure 2. The datasets for the experiment.](image)

3.2. Pre-processing data

3.2.1. Data cleansing. We use str() and summary() function to comprehend the data in the dataset. Because the data divided into five separate datasets, we must make sure that employee ID is a unique key and identical across all datasets.

3.2.2. Handling dates. Figure and Figure illustrate emp_intime and emp_outtime. To perform further analysis, we need to convert columns into an appropriate date and time format.

![Figure 3. Emp_intime data.](image) ![Figure 4. Emp_outtime data.](image)

After converting, we then calculate out-time and in-time to know actual working hours for every employee. Figure 5 shows the result of the calculation.
For our analysis purpose, we then categorized actual working hours into three categories: overtime (if working hours more than 8 hours), regular (if working hours between 7 and 8 hours), and early logout (if working hours less than 7 hours). Figure 6 illustrates the result.

3.2.3. Dealing with unknown values. Figure 7 illustrates the number of missing values. We found that 2.5 percent of data has missing values. We treat this by using the median of each column to replace the missing value. We also remove some variables which have only a single type of value.

3.3. Exploratory data analysis
In this step, we explored the data by numerical and non-numerical variables. For non-numerical variables, we used the bar chart to see the existence of attrition in each variable. Illustrates all non-numerical variables. From the illustration, we can see that the department, education field, working hours category, and work-life balance are variables that have more than 25 percent attrition.

3.4. Model selection and training
In this step, we removed outliers’ value from variables that exist in the data set. The target value of a variable (attrition) converted from “No”/“Yes” character to factor with levels 0/1.
To prepare and apply a model to this dataset, we break down the dataset into two subsets. We allocated 70 percent of items to the Training set and 30 percent items to the Test Set. In building the model, we remove multicollinearity through the VIF check. Shows 25 models created with refining based on AIC, residual deviance, and p-value. Our final model has eleven significant variables, as can be seen in Error! Reference source not found.

3.5. Model test and evaluation
Figure 12 shows a summary of the confusion matrix of the prediction model. Accuracy, sensitivity, and specificity of the prediction model is 75 percent, 73 percent, and 75 percent, respectively.

![Figure 5. Actual working hours.](image)

![Figure 6. Employee working hour’s categories.](image)

![Figure 7. Percent missing data.](image)
a) business travel, b) marital status, c) gender, d) department, e) education field, f) environment satisfaction, g) working hours category, h) education, i) work-life balance, j) job involvement, k) job role, l) job satisfaction, m) performance rating

**Figure 8.** Non-numerical variables bar chart.
4. Conclusion
From this study, we found that eleven variables that have a significant impact on employee attrition. The key driving factors are ‘a number of companies worked,’ ‘total working years,’ ‘years with current manager,’ ‘frequent business travel,’ ‘low environment satisfaction,’ ‘department human resources,’ ‘marital status – divorce,’ ‘marital status – married,’ ‘low job satisfaction,’ ‘early logout,’ and ‘overtime.’ Marital status is one factor that has an anomaly. If we see from the data, an employee with a single marital status has a more significant number in attrition than those who divorced and married. Two factors can drive Employee attrition. The first factor is drive by the employee factor itself, and the other is the company factor. Some employees, especially “junior” employees still want to have more experience. An employee with a small number of working years and companies worked has a more significant probability of attrition. To reduce employee attrition rate, the company needs to improve the human resource department by evaluating the working environment, job satisfaction, employee workload, and interaction between manager and employee.
References

[1] Mishra S N, Lama D R and Pal Y 2016 Human Resource Predictive Analytics (HRPA) for HR Management in Organizations *International Journal Of Scientific & Technology Research* 5(5) 33-35

[2] Marler J H and Boudreau J W 2016 An evidence-based review of HR Analytics *The International Journal of Human Resource Management* 28(1) 3-26

[3] Hoffman M and Tadelis S 2018 People Management Skills, Employee Attrition, and Manager Rewards: An Empirical Analysis *National Bureau of Economic Research*

[4] Frye A, Boomhower C, Smith M, Vitovsky L and Fabricant S 2018 Employee Attrition: What Makes an Employee Quit? *MU Data Science Review* 1(1)

[5] Raja D V A J and Kumar R A R 2016 A Study To Reduce Employee Attrition in IT Industries *International Journal of Marketing and Human Resource Management (IJMHRM)* 7(1) 1-14

[6] Bindra H, Sehgal K and Jain R 2019 Optimisation of C5.0 Using Association Rules and Prediction of Employee Attrition in *International Conference on Innovative Computing and Communications. Lecture Notes in Networks and Systems, Singapore*

[7] Mishra S N and Lama D R 2016 A decision making model for human resource management in organizations using data mining and predictive analytics *International Journal of Computer Science and Information Security* 14(5) 217

[8] Hosmer Jr D W, Lemeshow S and Sturdivant R X 2013 *Applied logistic regression* (New Jersey: John Wiley & Sons)

[9] Villar-Rubio E, Delgado-Alaminos J and Barrilao-González P 2015 Job satisfaction among Spanish tax administration employees: A logistic regression analysis *Journal of Labor Research* 36(2) 210-223

[10] Karhula K, Puttonen S, Ropponen A, Koskinen A, Ojajärvi A, Kivimäki M and Härmä M 2017 Objective working hour characteristics and work–life conflict among hospital employees in the Finnish public sector study *Chronobiology international* 34(7) 876-885

[11] Lohela-Karlsson M, Hagberg J and Bergström G 2015 Production loss among employees perceiving work environment problems *International archives of occupational and environmental health* 88(6) 769-777

[12] Halkos G and Bousinakis D 2017 The effect of stress and dissatisfaction on employees during crisis *Economic Analysis and Policy* 55 25-34

[13] Merriman K K, Sen S, Felo A J and Litzky B E 2016 Employees and sustainability: the role of incentives *Journal of Managerial Psychology* 31(4) 820-836