Proceedings of the 2023 International Conference on Management Research and Economic Development
DOI: 10.54254/2754-1169/20/20230198

Factors of Employee Attrition: A Logistic Regression Approach

Bingzhe Chen[1,a,*]

[1]*Tsinghua University, Beijing, 10010, China*
a. chenbz20@mails.tsinghua.edu.cn
**corresponding author*

Abstract: The problem of employee turnover has gradually become a common problem faced by companies around the world. Because it has several major negative impacts on the company's cultural and economic levels, it's urgent to solve this problem. In this paper, IBM's employee data is visualized to get the factors that affect employee attrition. At the same time, a predictive model based on logistic regression analysis was constructed to provide quantitative data for the impact of various factors on employee turnover. After multiple model evaluations, the goodness of fit and accuracy of the forecasting model were judged to be good. According to the data visualization and prediction model, the company's employee turnover factors were summarized, including both human resource factors and contextual factors such as age, salary, marital status. These influencing factors and the analysis of the relationship between them can be used as the theoretical basis for the company's future employee policies, providing direction and solutions for its work on reducing employee attrition.

Keywords: employee attrition, logistic regression model, data visualization

1. Introduction

1.1. Background

Employee attrition has been a problem for most companies and organizations around the world for a long time [1]. Especially nowadays, with the improvement of the average level of living standard and requirements in the society, a higher turnover rate has become a common phenomenon in many companies. According to recent statistics, the overall employee turnover rate in 2021 is as high as 53.7%, and the rate in many industries is nearly 19%, which is well above 10%, the basic standard for this data [2].

The reason why employee turnover needs to be taken seriously is that its harm and negative impact on the whole organization are immeasurable. The resignation of an outstanding employee will not only reduce the competitiveness of the company, but also mean that the energy and financial resources invested in the selection and training of this employee are wasted. At the same time, the resources devoted to selecting and developing the next employee to fill the vacancy will also take a large amount of the company's inherent resources [2].

DOI: 10.54254/2754-1169/20/20230198

In order to come up with a solution to the employee attrition problem and help organizations come up with better policies to retain employees with great performance and ability, the factors that lead to employees' willingness to leave need to be discovered.

1.2. Related Research

Factors found to be influencing attrition rate have been explored in several studies. Studies such as that conducted by Barpanda and Athira have shown that factors for IT sector can be divided into 2 categories: Human resource factors and contextual factors [3]. J.L.E.E.Liu had concluded that personal factors such as salary, job recognition and career development opportunities may be the key factors affecting employee turnover, after analyzing the personal information of 112 employees in a market [4]. G. V. Sridhar discussed the impact of objective factors such as job mobility, challenge and role importance on employee turnover on the basis of previous research [5]. However, most studies in employee attrition factors have only been carried out in a small number of areas, since for companies with different natures and fields, the types and psychology of their employees are different, thus the reasons for their resignation are not exactly the same. Therefore, the research on influencing factors must be targeted and specific, based on the company's own data, and then get the employee policies applicable to the organization.

It is also an effective solution to construct a prediction model for employee turnover. Jain, P.K., Jain, M. & Pamula, R. built a model based on machine learning to predict the probability of employee turnover using human resource dataset [6]. Raza A, Munir K, Almutairi M, Younas F, Fareed MMS. also construct a predicting model using machine learning techniques SVM and LR, based on limited amounts of factors [7]. Barpanda, Saswat, and Athira S. used a triangulation approach to understand the reasons for employee attrition including several human resource factors [3]. However, there has been little quantitative analysis of employee attrition using logistic regression model, which can be of great help while studying binary problems.

1.3. Objection

In this paper, the author managed to analyze the employee information and turnover rate of an IT company and used the method of data visualization to visually show which factors affect employee turnover and the degree of influence. At the same time, a prediction model was constructed to quantitatively analyze the importance of each factor's role in the employee turnover problem. Based on the above analysis, the author tried to provide data support for the company to eliminate employee turnover.

2. Methodology

2.1. Source of Data

The dataset used in this paper is from Kaggle, its initial sample consists of 14710 observations and 13 variables, 1 of whom is the target variable, which is the one named attrition [8]. Participants were all employees from the company IBM in 2021, covering attrition and non-attrition. The dataset contains several either numerical or categorical employee attributes, which can be used to construct a prediction on attrition. Table 1 shows the distribution of variables in the dataset and their descriptions.

DOI: 10.54254/2754-1169/20/20230198

Table 1: Information of the dataset.

Variable	Category	Description
Age	Numerical	Age of employees
Department	Categorical	Department of work
Distance from home	Numerical	Distance between employee's home and company
Education	Categorical	1-Below College; 2-College; 3-Bachelor; 4-Master; 5-Doctor
Education Field	Categorical	Employee's educational field
Environment Satisfaction	Categorical	1-Low; 2-Medium; 3-High; 4-Very High
Job Satisfaction	Categorical	1-Low; 2-Medium; 3-High; 4-Very High
Marital Status	Categorical	Employee's marital status
Monthly Income	Numerical	Employee's monthly income
Num Companies	Numerical	Number of companies worked prior to IBM
Work Life Balance	Categorical	1-Bad; 2-Good; 3-Better; 4-Best
Years At Company	Numerical	Current years of service in IBM
Attrition	Categorical	Employee attrition status(0 or 1)

2.2. Data Processing

(1) Categorical Variables

Stacked bar plots are capable of comparing subcategories within a large category and displaying the frequency ratio of each subcategory. In descriptive analysis of correlation coefficients between categorical variables, stacked bar plot is particularly useful. The dependent variable Attrition in the dataset is binary while several independent variables are also categorical, thereby, bar plot analyses were carried out using RStudio.

(2) Numerical Variables

The use of boxplot has a relatively long tradition within analyses that compare distribution of several groups. The basic relationship between binary dependent variable and numerical independent variables were collected using boxplots via RStudio.

(3) Correlation

The use of correlation coefficient is a well-established approach in measuring the degree of correlation between variables. There are three main types of correlation coefficients: Pearson, Spearman and Kendall. Criteria for selecting the subjects were as follows: Firstly, the calculation of the correlation coefficient needs to be applicable to both categorical and numerical variables; Secondly, the formula can still be applied when linear relationship between variables is not satisfied [9]. Considering the dataset used in this paper, Spearman correlation coefficient can be more useful for analysis. Spearman Correlogram was plotted using RStudio, the basic form of its formula is:

$$\rho_s = \frac{\sum_{i=1}^{N}(R_i-\bar{R})(S_i-\bar{S})}{\sqrt{\sum_{i=1}^{N}(R_i-\bar{R})^2(S_i-\bar{S})^2}} \quad (1)$$

DOI: 10.54254/2754-1169/20/20230198

In the equation above, R_i and S_i refer to the value level of the sample observations, $\bar{R}$ and $\bar{S}$ refer to the mean level of the observed variable, N refers to the total number of observations in this calculation.

(4) Predicting Model

After selecting an appropriate method for building the predicting model, the process will be achieved using SPSS.

2.3. Models

In an attempt to assess whether and how attrition is influenced by these employee attributes, a predicting model was established. Logistic regression approach was used to capture the complexities of the situation. The benefit of this approach is that it is a probabilistic nonlinear regression model that predicts and judges the probability of an outcome, which allows all categories of independent variables. In practical application, the logistic regression model is already a sophisticated approach to predict dependent variables, especially binary variables [10].

The basic form of binary logistic regression is:

$$logit(P) = \ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \tag{2}$$

In the equation above, the term 'X' refers to p independent variables, the term 'P' refers to the probability of the event occurring $P(Y = 1|X)$, β_0 is a constant term, β_1, β_2, ⋯, β_i can broadly be defined as the partial regression coefficient. $\frac{P}{1-P}$ is generally understood to mean odd ratio(OR), it describes the degree to which the covariable is associated with the outcome when other factors are controlled. Further derivation of this equation yielded a non-linear function of multiple independent variables:

$$P = \frac{\exp\left(\beta_0 + \sum_{i=1}^{p} \beta_i X_i\right)}{1 + \exp\left(\beta_0 + \sum_{i=1}^{p} \beta_i X_i\right)} \tag{3}$$

Substituting all observed values $(x_i\ ,\ y_i)$, $i = 1,2,\cdots,n$ and parameters to be estimated β, the following equations can be obtained:

$$\sum_{i=1}^{n} [y_i - P(x_i)] = 0 \tag{4}$$

$$\sum_{i=1}^{n} x_{ij} [y_i - P(x_i)] = 0 \tag{5}$$

In the equation above, $j = 1,2,\cdots,p$. Using variance combined with maximum likelihood estimation, the likelihood equation of the second-order partial derivative was obtained from the perspective of the matrix:

$$\frac{\partial^2 L(\beta)}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 P_i (1 - P_i) \tag{6}$$

$$\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_l} = -\sum_{i=1}^{n} x_{ij} x_{il} P_i (1 - P_i) \tag{7}$$

DOI: 10.54254/2754-1169/20/20230198

In the equation above, $j, l = 0,1,2,\cdots,p$, the term P_i refers to $P(X_i)$. Estimation of coefficients can be made using their variance and covariance together with a (p+1)-order matrix of observation information, but they will be inversely proportional. Another approach where $Var(\beta_j)$ and $Cov(\beta_j, \beta_l)$ are used was chosen to conduct this estimation:

$$\widehat{SE}(\widehat{\beta_j}) = [\widehat{Var}(\widehat{\beta_j})]^{1/2} \tag{8}$$

2.4. Data Processing Evaluation Index

(1) Model overall evaluation

To see if the model is of value, Omnibus Test of Model Coefficients was carried out. The test output the likelihood ratio test of whether all parameters in the model equal to 0. $P < 0.05$ means that in the fitted model, the OR value of at least one of the included variables is statistically significant, so that the model is also of significance.

(2) Goodness of fit

Several methods currently exist for the measurement of goodness of fit: Pearson χ^2, Deviance and Hosmer-Lemeshow goodness of fit. The first two methods are particularly useful in studying models that do not involve continuous independent variables. When continuous variables exist, some covariates have too many different values, resulting in a large number of covariate types [11]. So, Hosmer-Lemeshow was adopted in this paper, using the following statistical equation:

$$HL = \sum_{g=1}^{G} \frac{y_g - n_g\widehat{p_g}}{n_g\widehat{p_g}(1-\widehat{p_g})} \tag{9}$$

In the equation above, the term $'G'$ refers to number of groups $G \leq 10$; n_g refers to number of cases in group g; y_g refers to number of observations for the gth group of events; $\widehat{p_g}$ refers to probability of predicted event for group g; $n_g\widehat{p_g}$ refers to predicted number of events, it equals to the sum of the predicted probabilities for the gth group.

(3) Model predictive power

To see the predictive power of the model, percentage accuracy in classification was measured, which shows the relationship between the observed value and the predicted value. In a binary classification problem, the predicted results are divided into 4 types: True Positive, False Positive, True Negative and False Negative. "TP", "FP", "TN" and "FN" respectively refer to the number of samples corresponding to these 4 results. The formula for calculating accuracy is:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{10}$$

3. Results and Discussion

3.1. Data Visualization

3.1.1. Categorical Variables

Percent stacked bar plots below show whether each categorical attribute has an influence on employee's attrition status. As shown in figure 1(a), the sales department has significantly higher probabilities losing employees than the other two departments, while the research & development department has the lowest probabilities. figure 1(b) presents that with great environment satisfaction comes low probabilities of attrition. The same result applies to the relationship between job satisfaction and attrition. It can be seen from figure 1(d) that the attrition rate of single employees is higher than that of divorced and married employees. figure 1(e) shows that

DOI: 10.54254/2754-1169/20/20230198

employees with better life-work balance are less likely to leave the company. From figure 1(f) it can be seen that the higher the education level of employees, the lower the attrition rate.

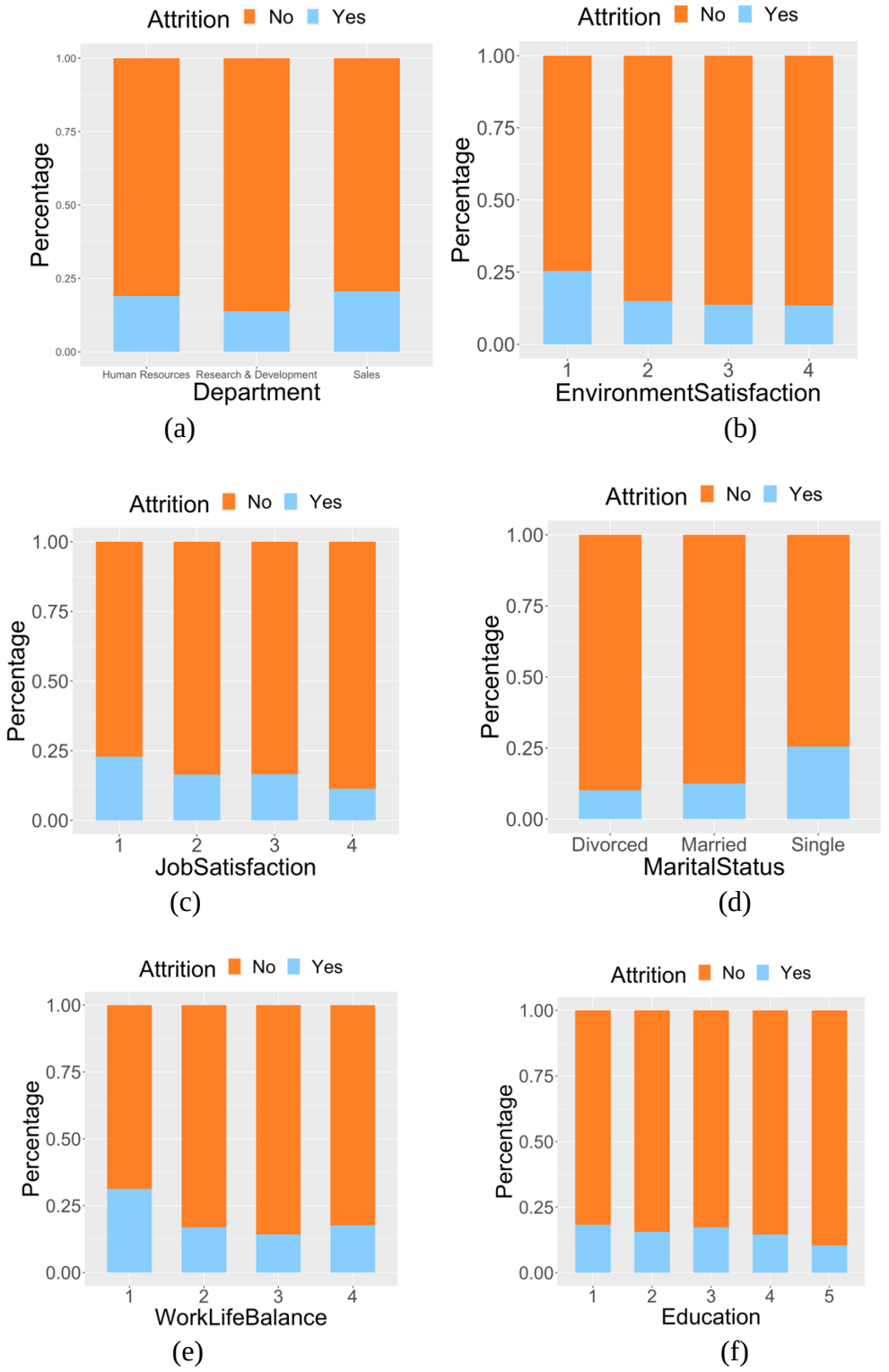


Figure 1: Percent stacked bar plots of IBM employees' information. (Photo credit: Original)

DOI: 10.54254/2754-1169/20/20230198

Due to the large number of variable categories in the education field, a grouped bar plot was plotted for further analysis. From figure 2 it can be seen that the number of employees who were lost is much lower than the number of employees who were not, no matter which field the employee was educated in. A comparison of the six results reveals that employees majored in life sciences have the most attrition quantity; employees majored in medical have the second highest one; marketing, technical degree, human resources and others are education fields that have the lowest attrition quantity.

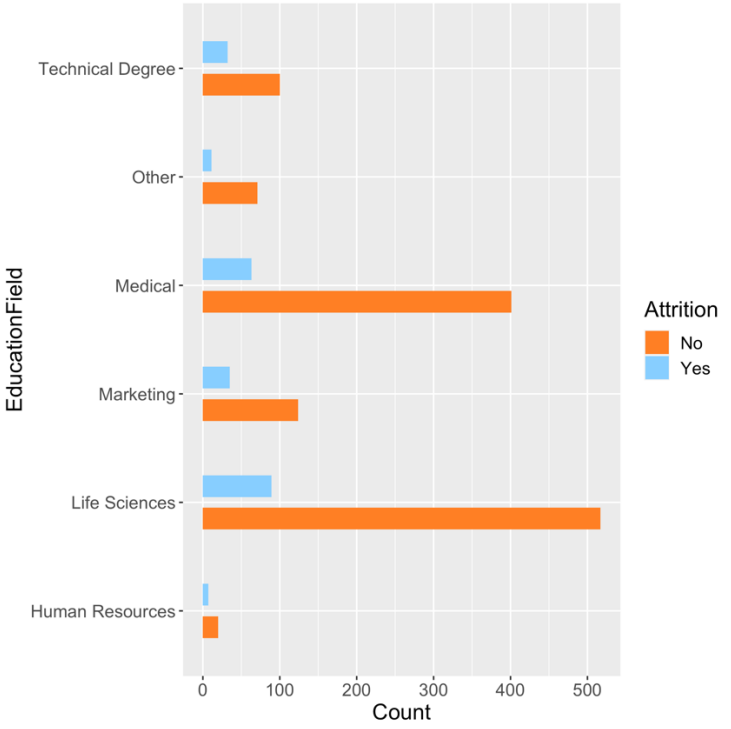
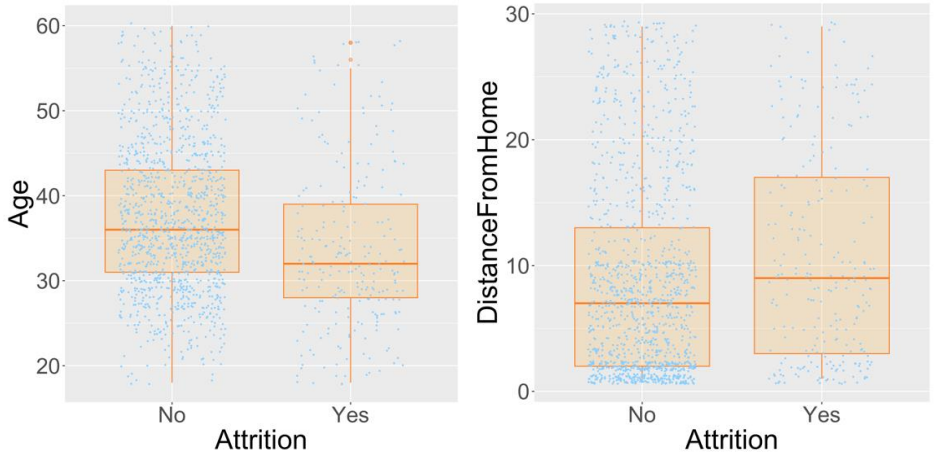


Figure 2: Grouped bar plot of education field. (Photo credit: Original)

3.1.2. Numerical Variables

The results of the boxplot analysis are set out in figure 3.



(a) (b)

DOI: 10.54254/2754-1169/20/20230198

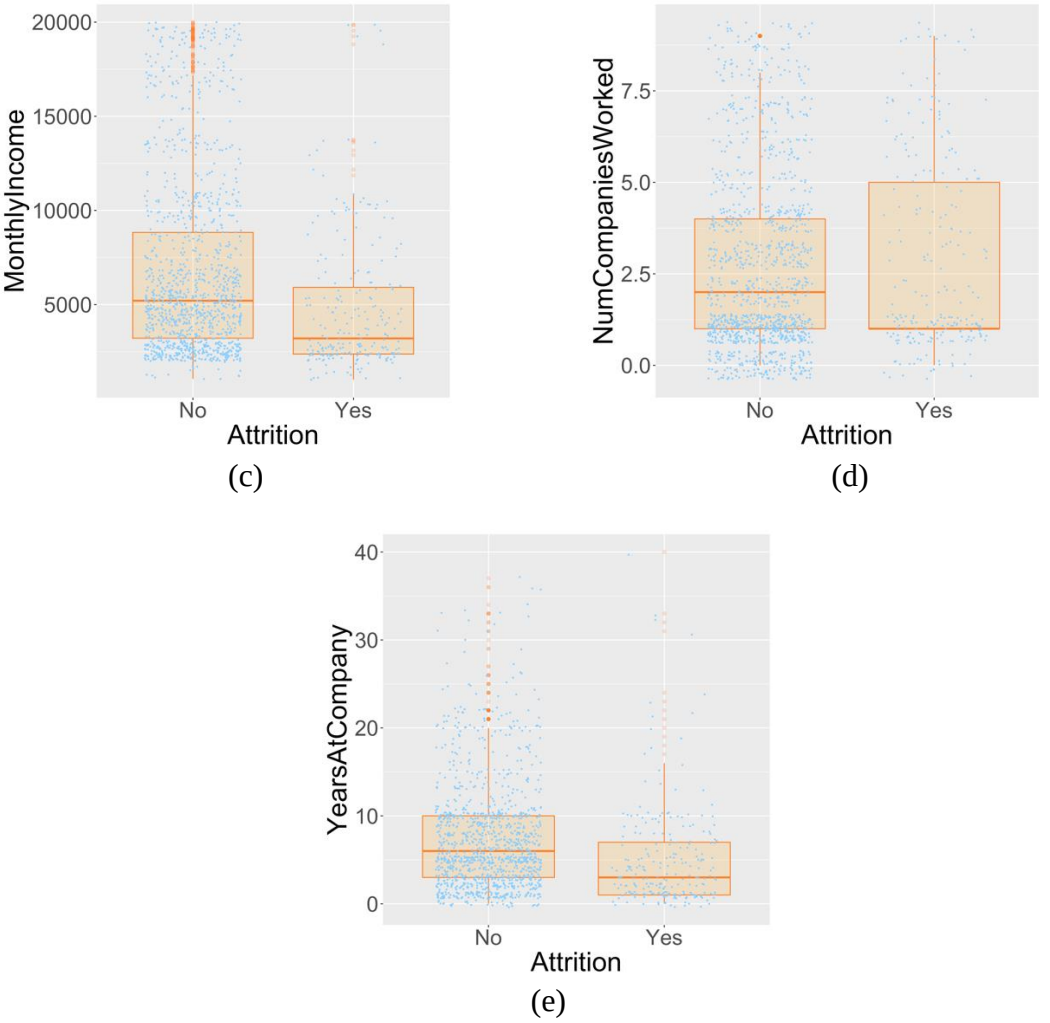


Figure 3: Boxplots of IBM employees' information. (Photo credit: Original)

From the graph it can be seen that the number of employees who were lost is much lower than the number of employees who were not; The mean value of lost employee's age is about 5 years lower than the one of stayed employee's; Employee whose home is far from the company is more likely to leave the company; Monthly income has a positive influence on employee's staying, stayed employees' mean income is significantly higher than lost employees'. On average, number of companies where employees had worked before were shown to have impact on attrition: The more companies one had been in, the less likely he/she would leave his/her current position. figure 3(e) illustrates that employees with longer years at this company are less likely to leave.

3.1.3. Correlation

Figure 4 below shows the correlation between these numerical variables. When the correlation value is more than 0.7, generally it means there is a strong relationship between the two variables. According to the graph, most of the variables are independent from each other. There might be a slight connection between age and income, and another slight connection between years at company and income, but their relationship is not close enough to have a negative impact on the predicting model. As a result, no variable would be filtered during the predicting process.

DOI: 10.54254/2754-1169/20/20230198

Figure 4: Spearman correlogram. (Photo credit: Original)

3.2. Statistical Analysis

The results obtained from the Omnibus test of model coefficients are presented in table 2.

Table 2: Omnibus tests of model coefficients.

		Chi-square	df	Sig.
Step 1	**Step**	187.632	18	< .001
	Block	187.632	18	< .001
	Model	187.632	18	< .001

$P < 0.001$ suggests that in the fitted model, the OR value of at least one of the independent variables is statistically significant, which means the model was constructed successfully.

Table 3: Hosmer and Lemeshow test.

Step	Chi-square	df	Sig.
1	11.362	8	.182

Table 3 shows the result of the Hosmer-Lemeshow test. $P = 0.182 > 0.05$ means that the information in the data has been fully extracted, and the model's goodness of fit conforms to the standard.

DOI: 10.54254/2754-1169/20/20230198

Table 4: Classification table.

	Observed		Predicted		
			Attrition		Percentage Correct
			No	Yes	
Step1	Attrition	No	1226	7	99.4
		Yes	201	36	15.2
	Overall Percentage				85.9

As can be seen from table 4, the model can correctly classify 85.9% of the observations, thus, it has a high prediction accuracy. Among them, the prediction accuracy rate for employees who have not been lost is as high as 99.4%.

Table 5: Variables in the equation.

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1	**Age**	-.035	.011	10.721	1	.001	.966
	Department			6.465	2	.039	
	Department(1)	-.352	.540	.423	1	.515	.704
	Department(2)	-.492	.194	6.464	1	.011	.611
	DistanceFromHome	.031	.009	11.922	1	< .001	1.032
	Education	-.028	.078	.128	1	.720	.973
	EducationField			11.475	5	.043	
	EducationField(1)	.313	.729	.184	1	.668	1.367
	EducationField(2)	-.615	.256	5.755	1	.016	.541
	EducationField(3)	-.263	.336	.610	1	.435	.769
	EducationField(4)	-.778	.268	8.460	1	.004	.459
	EducationField(5)	-.650	.406	2.563	1	.109	.522
	EnvironmentSatisfaction	-.297	.070	17.933	1	< .001	.743
	JobSatisfaction	-.287	.069	17.363	1	< .001	.750
	MaritalStatus			38.476	2	< .001	
	MaritalStatus(1)	-1.150	.226	25.784	1	< .001	.317
	MaritalStatus(2)	-.876	.170	26.725	1	< .001	.416
	MonthlyIncome	.000	.000	13.145	1	< .001	1.000
	NumCompaniesWorked	.116	.032	13.212	1	< .001	1.123
	WorLifeBalance	-.282	.106	7.101	1	.008	.754
	YearsAtCompany	-.022	.019	1.295	1	.255	.978
	Constant	3.501	.614	32.467	1	< .001	33.154

The results, as shown in table 5, indicate which independent variable has statistical significance in this model, together with the correlation between significant variables and the dependent variable. The column 'Sig.' presents the value of P for each independent variable, 'P<0.05' means that the variable is of statistical significance. The column 'Exp(B)' presents the OR value of each variable, which reveals that: For categorical or discrete variables, how many times more likely are subjects with high ratings to have the outcome, compared to subjects with lower ratings; For

DOI: 10.54254/2754-1169/20/20230198

continuous numerical variables, how many times the likelihood of the outcome will increase by, for each unit increase in the independent variable.

It can be seen from the data in table 5 that except for education level and years at company, all independent variables are statistically significant in this model. Educational field, number of companies worked before, distance from home and monthly income have OR values over 1, which means their impact on whether employees are attrition is significant. Other variables like age, work-life balance, environment satisfaction and job satisfaction also have a relatively minor impact on the outcome due to the fact that their OR values are between 0.5 and 1. The rest of variables' influence on the results reflected in this predicting model is relatively slight.

3.3. Limitation

As a basic model, logistic regression model is simple and intuitive, but it also has many limitations. This model is unable to perform intersection and feature screening. It is a weak classifier and has limitations in adaptability to data and scenarios. Also the logistic regression model's learning ability is not as strong as the algorithm decision tree's. At the same time, the simulation model of logistic regression is incomplete and can only be used to predict classification results until all related independent variables are determined. Apparently, the relevant independent variables are not all covered in this dataset, variables like gender, performance rating, job level, overtime are all reasonably related to attrition while not being taken into consideration in this dataset.

4. Conclusion

On the basis of data visualization and logistic regression prediction model, this study reveals the main factors that cause the employee turnover problem of the sample company IBM, as well as the correlation between these factors and the degree of influence on the results.

a. The company's employee turnover problem is the result of the employee's own situation and the company's objective situation. For the employee's own situation such as environmental satisfaction, job, satisfaction, age, marriage, education level and home address all have significant impact on employee attrition. Objective factors like salary, department category, seniority also play important roles in employee attrition.

b. The binary classification prediction model based on logistic regression is suitable for the prediction of employee turnover. It can take into account both categorical and numerical variables along with the correlation between variables. The model has excellent performance in terms of goodness of fit and prediction accuracy and can provide valuable prediction results.

c. The simulation results of the predictive model explain several factors that have the greatest impact on employee turnover in this company: educational field, number of companies worked before, distance from home and monthly income. This provides direction and ideas for the company's future policy reformation.

Further work will extend this research to analysis for more comprehensive human resource factors and contextual factors, so that the prediction model has better goodness of fit and accuracy. This will make it more applicable to a wider range of practical applications.

References

[1] *Qutub A, Al-Mehmadi A, Al-Hssan M, Aljohani R, Alghamdi HS 2021 Prediction of employee attrition using machine learning and ensemble methods Int. J. Mach. Learn. Comput. 11(2) 110-4.*

[2] *Here's What Your Turnover and Retention Rates Should Look Like. 15 June 2021. Available online: https://www.ceridian.com/blog/turnover-and-retention-rates-benchmark (accessed on 2 Feb 2023).*

[3] *Barpanda S, Athira S 2022 Cause of Attrition in an Information Technology-Enabled Services Company: A Triangulation Approach International Journal of Human Capital and Information Technology Professionals*

DOI: 10.54254/2754-1169/20/20230198

(IJHCITP) 13(1) 1-22.

[4] Lee Liu J 2014 Main causes of voluntary employee turnover a study of factors and their relationship with expectations and preferences PhD thesis (Chile: Univ. Chile).

[5] Sridhar GV, Venugopal S, Vetrivel S 2018 Employee Attrition and Employee Retention-Challenges & Suggestions Conf. on Economic Transformation with Inclusive Growth-2018 (Chennai) vol 1 p 16.

[6] Jain PK, Jain M, Pamula R 2020 Explaining and predicting employees' attrition: a machine learning approach SN Appl. Sci. 2 1-11.

[7] Raza A, Munir K, Almutairi M, Younas F, Fareed MM 2022 Predicting Employee Attrition Using Machine Learning Approaches. Appl. Sci. 12(13) 6424.

[8] Kaggle. Employee-Attrition-Rate. Available online: https://www.kaggle.com/datasets/prachi13/employeeattritionrate

[9] Zhang SQ, Lv JN, Jiang Z, Zhang L 2009 Study of the Correlation Coefficients in Mathematical Statistics Mathematics in Practice and Theory 39(19) 102-7.

[10] Wang QQ, Yu SC, Qi X, Hu YH, Zheng WJ, Shi JX, Yao HY 2019 Overview of logistic regression model analysis and application Zhonghua yu fang yi xue za zhi [Chinese journal of preventive medicine] 53(9) 955-60.

[11] Yazici B, Alpu Ö, Yang Y 2007 Comparison of goodness-of-fit measures in probit regression model Communications in Statistics—Simulation and Computation®. 36(5) 1061-73.