Job Satisfaction Prediction and Machine Learning Technique

Youngkeun Choi (penking1@smu.ac.kr)
Sangmyung University
Jae Choi
University of Texas at Dallas

Research Article

Keywords: Machine learning, Data science, Generalized linear model, Human resource management, Job satisfaction

Posted Date: May 27th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1683972/v1

Abstract

This paper aims to develop a reliable job satisfaction prediction model using machine learning technique. For this, this study used the dataset which is available at IBM Watson Analytics and applied generalized linear model including linear regression and binomial classification. This study essentially had two primary approaches. Firstly, this paper intends to understand the role of variables in job satisfaction prediction modeling better. Secondly, the study seeks to evaluate the predictive performance of generalized linear model including linear regression and binomial classification. In these results, first, we can predict employees' job satisfaction with a lot of individual factors. Second, for each model, our model showed the outstanding predictive performance. The pre-access and modeling methodology used in this paper can be viewed as a roadmap for the reader to follow the steps taken in this study and to apply procedures to identify the causes of many other human resource management problems.

1. Introduction

Computer and internet-based technology development is a source of a rapid increase in the amount and availability of data worldwide. We can get larger-scale data more easily than ever, allowing for insights in the form of new information processing or decision-making tools through analytical formulas and rule-development possibilities (data-processing algorithms) to solve problems.

Most recently, the spread of intelligent machine learning algorithms in the field of computer science has developed a powerful quantitative method that elicits insights from industrial data. Supervised machine learning methods (an analysis of data sets labeled on computers learned in many past years) include biology and medical science (Bakry et al., 2016), transportation (Mathisa & Ragusa, 2016), political science (Durant & Smith, 2006), and many other areas. In response to advances in information technology, researchers studied many machine learning approaches to improve HR (human resources) management outcomes (Li et al., 2011; LaFayette et al., 2019; Hughes et al., 2019).

Big data has become a popular label for many data analytics efforts. The original term Big Data emerged to define a technological revolution that enabled massive data collection (Jacobs 2009). Since then, the term has moved to different domains to represent different aspects of the analysis, depending on the circumstances in which big data has been mentioned. The term is now used to represent data processing capabilities and data characteristics and includes both technical and commercial aspects of data collection activities (Nunan & Di Dominico 2017). Mayer-Schönberger and Cukier (2013) regard big data as new features that allow them to collect vast amounts of information and analyze it immediately (Kitchin 2014). In a similar vein, Boyd and Crawford (2012) suggests that big data does not necessarily have to be a statement describing the size of the data, but instead is a term for the ability to search, aggregate and cross-reference large data sets.

Human resources have enormous amounts of data. The system includes built-in data, such as employees, information, participation scores, and performance records. Every detail of an individual or organization, every aspect that can be done or documented, is lost immediately after use. As a result, organizations lose the ability to extract valuable information, perform detailed analysis, and provide new opportunities and benefits as well as knowledge. The customer's name and address, the purchase, and everything in the hands of the employee have become very important in everyday life. Therefore, data is a fundamental element of organizational success. Scope, transformation, and rapid change in this type of data require new types of big data analytics and a variety of analysis and storage methods. This absolute amount of big data needs to be correctly analyzed and relevant information removed. The HR department began to use data analytics to identify the highest performance ever used, improve withholding rates, and start to benefit from everyone's happy participation. HR experts quickly began to embrace data analytics. Now we think about the spread of information and information available today through the evolution of technology and the Internet. As storage capabilities and data classification methods grow, massive amounts of data are available. More and more data is generated every 1 second and stored and analyzed to extract values. Organizations also need to make the most of their vast amounts of stored data because of the low cost of storing data.

Organizations bear high costs right from recruitment, selection, induction, and training them to make them work according to the needs of the organization. Any crucial employee who quits the organization is a big loss for the organization not only in monetary terms but also that employee takes away a lot of knowledge, information, experience, and skills which are significant for the organizational growth but goes in the hands of the competitor firm with an employee who decides to leave. Therefore, retaining employees is retention is the critical driver for organizational success. There are numerous reasons for attrition. If the employees feel dissatisfied due to any reason, they leave the current employer and join where they get better opportunities to satisfy their personal and professional needs. This issue can be well understood by connecting it with job satisfaction. Job satisfaction represents a combination of positive or negative feelings that workers have towards their work. Meanwhile, when a worker employed in a business organization, it brings with it the needs, desires, and experiences which determine expectations that he has dismissed. Job satisfaction represents the extent to which expectations are and match the real awards. Job satisfaction is closely linked to that individual's behaviour in the workplace (Davis et al.,1985).

This paper describes the key machine learning algorithms used to address employees’ job satisfaction issues. A new contribution to this paper is to explore the application of machine learning. The preprocessing and modeling methodology used in this paper can be viewed as a roadmap for the reader to follow the steps taken in this study and to apply procedures to identify the causes of many other HR problems. Therefore, this paper provides a quick, immediate, and easy way to select potential employees. A unique benefit can be provided to HR departments.

2. Related Study

Job satisfaction is a worker’s sense of achievement and success on the job (Cherif, 2020; García et al., 2019). It is generally perceived to be directly linked to productivity as well as to personal well-being. Job satisfaction implies doing a job one enjoys, doing it well, and being rewarded for one's efforts. Job satisfaction further means enthusiasm and happiness with one's work. Job satisfaction is the key ingredient that leads to recognition, income, promotion, and the achievement of other goals that lead to a feeling of fulfillment (Kaliski,2007). Job satisfaction can also be defined as the extent to which a worker is a content with the rewards he or she gets out of his ore her job, particularly in terms of intrinsic motivation (Statt, 2004).

The term job satisfaction refers to the attitudes and feelings people have about their work. Positive and favorable attitudes toward the job indicate job satisfaction. Negative and unfavorable attitudes toward the job indicate job dissatisfaction (Armstrong, 2006). Job satisfaction is the collection of feelings and beliefs that people have about their current job. People's levels of degrees of job satisfaction can range from extreme satisfaction to extreme dissatisfaction, in addition to having attitudes about their jobs as a whole. People also can have attitudes about various aspects of their jobs, such as the kind of work they do, their coworkers, supervisors or subordinates, and their pay (George et al., 2008).

Job satisfaction is a complex and multifaceted concept that can mean different things to different people. Job satisfaction is usually linked with motivation, but the nature of this relationship is not clear. Satisfaction is not the same as motivation. Job satisfaction is more of an attitude, an internal state. It could, for example, be associated with a personal feeling of achievement, either quantitative or qualitative (Mullins, 2005).

We consider that job satisfaction represents a feeling that appears as a result of the perception that the job enables the material and psychological needs (Aziri, 2008). Job satisfaction can be considered as one of the main factors when it comes to the efficiency and effectiveness of business organizations. The new managerial paradigm, which insists that employees should be treated and considered primarily as human beans that have their wants, needs, personal desires, is a very good indicator of the importance of job satisfaction in contemporary companies. When analyzing job satisfaction, the logic that a satisfied employee is a happy employee and a happy employee is a successful employee.

The importance of job satisfaction specially emerges to surface if had in mind the many negative consequences of job dissatisfaction such a lack of loyalty, increased absenteeism, increase number of accidents, etc. Spector (1997) lists three important features of job satisfaction. First, organizations should be guided by human values. Such organizations will be oriented towards treating workers fairly and with respect. In such cases, the assessment of job satisfaction may serve as a good indicator of employee effectiveness. High levels of job satisfaction may be a sign of a good emotional and mental state of employees. Second, the behavior of workers, depending on their level of job satisfaction, will affect the functioning and activities of the organization's business. From this, it can be concluded that job satisfaction will result in positive behavior and vice versa, dissatisfaction from work will result in negative behavior of employees. Third, job satisfaction may serve as indicators of organizational activities. Through job satisfaction evaluation, different levels of satisfaction in different organizational units can be defined. Still, in turn, it can help as a good indication regarding which organizational unit changes that would boost performance should be made.

One way for organizations to solve this problem is to use machine-learning technology to predict employees' job satisfaction so that their leaders and HR can take proactive measures or plan succession for preservation. But the machine learning technology historically used to solve this problem does not account for data noise in most HR Information Systems (HRIS) (Mauro & Borges-Andrade, 2020). Most organizations have not prioritized investments in efficient HRIS solutions that capture employee data during their tenure. One of the key factors is a limited understanding of benefits and costs. Measuring the return on investment in HRIS remains difficult (Jahan, 2014). This causes noise in the data, which weakens the generalization of these algorithms. To illustrate this concern, we use an algorithm to predict employee's job satisfaction. As is common in these issues, machine-learning technology can create algorithms based on employee attributes that are relevant to the job performance of your current workforce. Despite the causal relationship between traits such as gender and job satisfaction, the algorithm for promoting more men is unreliable because job performance itself can be a characteristic of biased indicators, current workforce, and data. It can also be distorted by the way it was employed in the past (e.g., very few women are employed).

3. Methodology

3.1 Dataset

The dataset used in this paper was related to job satisfaction and is available at IBM Watson Analytics in IBM Community. The key to success in any organization is attracting and retaining top talent. We can be an HR analyst at my company, and one of our tasks is to determine which factors keep employees at my company and which prompt others to leave. We need to know what factors I can change to prevent the loss of good people. Watson Analytics is going to help. Watson Analytics has data about past and current employees in a spreadsheet on desktop. It has various data points on our employees, but Watson Analytics is most interested in whether they're still with my company or whether they've gone to work somewhere else. And Watson Analytics wants to understand how this relates to workforce attrition. For each of 10 years, it shows employees that are active and those that terminated. The intent is to see if individual terminations can be predicted from the data provided. To help with algorithmic development, the organizers provided the types of a data stream for a large set of individual factors. These variables are listed and defined in Table 1.

< Table 1 > The measurements of variables

Variables	Measurement
Age	Integer
Attrition	Binomial (True or False)
BusinessTravel	Polynomial (Travel_Rarely 71%, Travel_Frequently 19%, Other 10%)
DailyRate	Integer
Department	Polynomial (Research & Development 65%, Sales 30%, Other 4%)
DistanceFromHome	Integer
Education	Integer (1 'Below College' 2 'College' 3 'Bachelor' 4 'Master' 5 'Doctor'
EducationField	Polynomial (Life Sciences 41%, Medical 32%, Other 27%)
EmployeeCount	Integer
EmployeeNumber	Integer
EnvironmentSatisfaction	Integer
Gender	Binomial (Male 60%, Female 40%)
HourlyRate	Integer
JobInvolvement	Integer (1 'Low' 2 'Medium' 3 'High' 4 'Very High')
JobLevel	Integer
JobRole	Polynomial (Sales Executive 22%, Research Scientist 20%, Other 58%)
JobSatisfaction	Integer (1 'Low' 2 'Medium' 3 'High' 4 'Very High')
MartialStatus	Polynomial (Married 46%, Single 32%, Other 22%)
MonthlyIncome	Integer
MonthlyRate	Integer
NumCompniesWorked	Integer
Over18	Binomial (True or False)
OverTime	Binomial (True or False)
PercentSalarHike	Integer
PerformanceRating	Integer (1 'Low' 2 'Good' 3 'Excellent' 4 'Outstanding')
RelationshipSatisfaction	Integer (1 'Low' 2 'Medium' 3 'High' 4 'Very High')
StandardHours	Integer
StockOptionLevel	Integer

Variables	Measurement
TotalWorkingYears	Integer
TrainingTimesLastYear	Integer
WorkLifeBalance	Integer (1 'Bad' 2 'Good' 3 'Better' 4 'Best')
YearsAtCompany	Integer
YearsinCurrentRole	Integer
YearsSinceLastPromotion	Integer
YearsWithCurrManager	Integer

3.2 Generalized linear model

The generalized linear model (GLM) provides a very broad and popular family for statistical analysis. For a particular choice of GLM, a measure of the model's predictive power can be useful for evaluating the practical importance of the predictors and for comparing competing GLMs, for example, models with different link functions or with different linear predictors. In ordinary regression for normal response, the multiple correlation R, and the coefficient of determination R^2 serve this purpose. Many summary measures of predictive power have been proposed (Mittlbock & Schemper, 1996) for GLMs. We now describe three of the main types of these measures and their shortcomings. First, these statistics measure the association between the ordered values of the response outcomes and the fitted values. The most popular measure of this type is the concordance index (Harrell et al., 1982), denoted by c. Consider those pairs of observations that are untied on Y. The index c equals the proportion of such pairs for which the predictions $\hat{Y}$ and the outcomes Y are concordant, the observation with the larger Y also having the larger $\hat{Y}$. For a binary response, c is related to a widely used measure of diagnostic discrimination, the area under a receiver operating characteristic (ROC) curve (Harrell et al., 1982). Various software packages, including S-plus (Harrell et al., 1996), STATA and SAS (PROC LOGISTIC), report this measure. Appealing features of c are its simple structure and its generality of potential application. Because c utilizes ranking information only, however, it cannot distinguish between different link functions, linear predictors, or distributions of the random components that yield the same orderings of the fitted values. For a binary response with a single linear predictor, for instance, the concordance index c assumes the same value for logit and complementary log-log link functions, even though the models are quite different; as long as the predicted values remain monotonic, c also remains the same when polynomial terms are added to the linear predictor.

Second, in ordinary linear regression with the normal model assuming constant variance, the coefficient of determination, R^2, describes the proportion of variance in Y explained by the model. It has been applied to other types of responses. For binary outcomes, for instance, let denote the model-based ML estimate of the probability of a positive response for subject *i*, and let $\bar{y}$ denote the sample proportion of positive responses. The sample measure (Efron, 1978) is defined as:

$$R^2 = 1 - [\textstyle\sum_{i=1}^n (y_i - \hat{\pi}_i)^2]/[\sum_{i=1}^n (y_i - \bar{y})^2]$$

Some have criticized the use of R^2 for non-normal GLMs because of restrictions in possible values to the lower end of the usual [0; 1] scale and sensitivity to the prevalence of the outcome (Cox & Wermuth, 1992). However, others have argued that sensitivity to prevalence is a strength (Hilden 1991) that a model with a low value of R^2 may still be helpful for prediction (Ash, & Shwarta, 1999), and that R^2 captures information (Ash, & Shwarta, 1999) not reflected by *c*. For an arbitrary measure of variation D(.), a natural extension (Haberman, 1982) of R^2 takes the form

$$\frac{\sum_{i=1}^n D(Y_i) - \sum_{i=1}^n D(Y_i|X_i)}{\sum_{i=1}^n D(Y_i)}$$

where *D(Yi)* denotes the variation for the *i*th observation and *D(Yi/Xi)* denotes the variation for the *i*th observation given the fixed value *Xi* of *X*. For a binary response, the proposed variation functions include squared error, prediction error, entropy and linear error (Efron, 1978). For a categorical response, proposed variation functions include the Gini concentration measure and the entropy measure (Haberman, 1982). Variation measures have also been proposed for other variants of the usual continuous response, such as a variety of measures for censored responses in survival analysis (Korn, & Simon, 1990). Like *c*, an appealing aspect of measures based on variation functions is their simple structure, one that is well familiar to those who use R^2 for normal data. A disadvantage is that their numerical values can be difficult to interpret, depending on the choice of variation function. Although the measures may be useful in a comparative sense, many biostatisticians and most of the medical scientific community would find it difficult to envision what a 50 per cent reduction in entropy represents, for instance.

Third, let *l* denote the likelihood function and let *L* = log *l* denote the log-likelihood. Let LM = log *l* M denote the maximized log-likelihood under the model of interest. Let *L*S denote the maximized log-likelihood under the saturated model, which has as many parameters as observations, and let *L*0 denote the maximized log-likelihood under the null model, which has only an intercept term. Let *D*M = − 2(*L*M − *L*S) and *D*0 = − 2(*L*0 − *L*S) denote the deviances for the model of interest and the null model. A summary measure based on the likelihood function is (Theil, 1970).

3.3 Preprocessing and data mining models

Statistical and data mining techniques have been utilized to construct decision prediction models. The data mining techniques can be used to discover interesting patterns or relationships in the data, and predict or classify the behavior by fitting a model based on available data. In the case where the learning dataset and the test dataset are separated for machine learning, the test dataset must satisfy the following requirements. First, the training dataset and the test dataset must be created in the same format. Second, the test dataset should not be included in the training dataset. Third, the training dataset and the test dataset must be consistent in data. However, it is very difficult to create a test data set that meets these requirements. In data mining, various verification frameworks using one dataset have been developed to solve this problem. This study uses the split validation operator provided by RapidMiner to support this. The operator splits the input dataset into a training dataset and a test dataset to support performance evaluation. This study selects relative segmentation among the segmentation method parameters of this operator and uses 70% of input data as learning data.

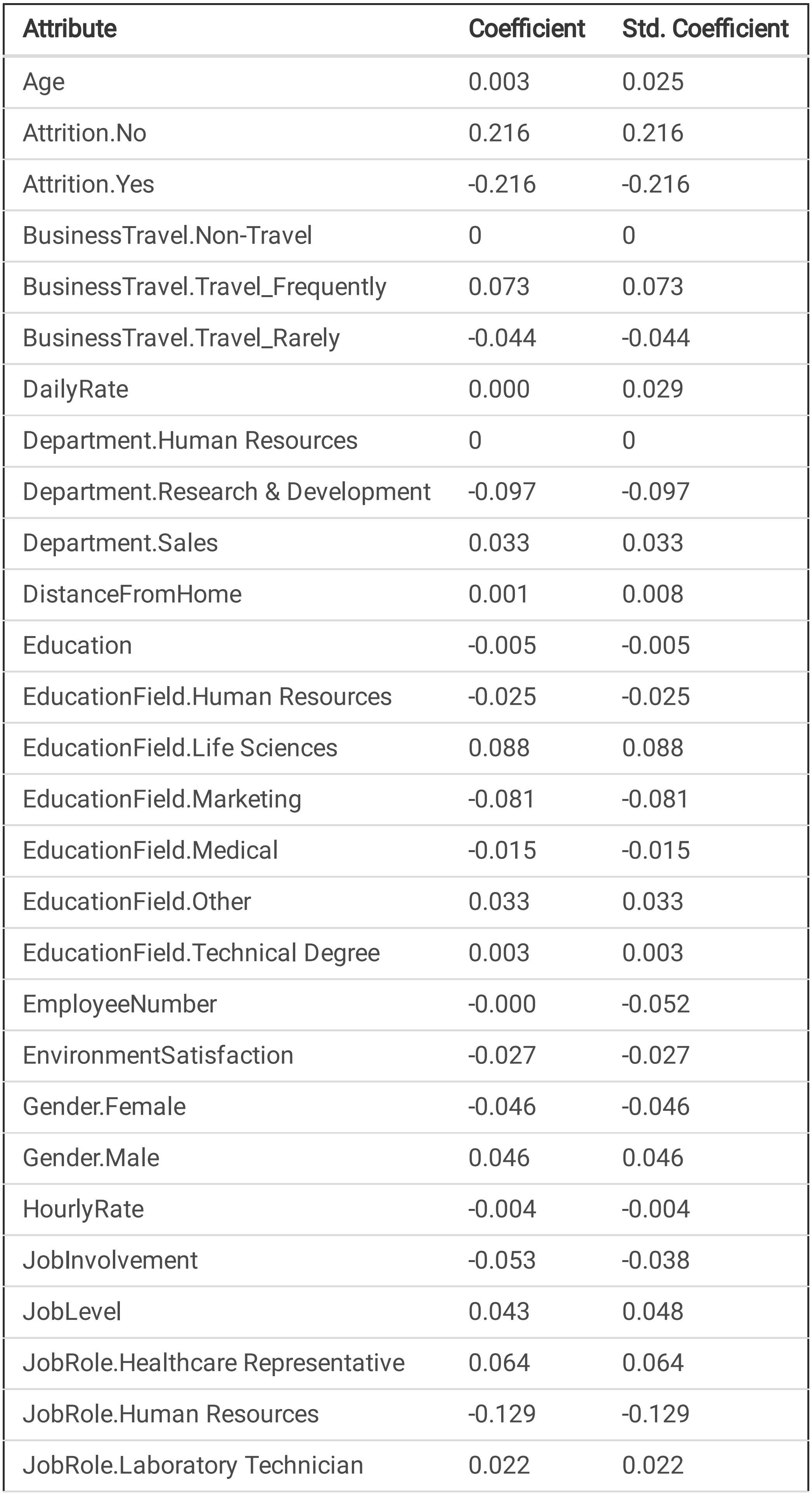
4. Results

In this study, we want to analyze the factors in the effect on job satisfaction. The job extension has a range of 1 to 4. The purpose of this analysis is to examine whether the GLM can address two types of problems: numerical prediction, binomial classification. Therefore, the numerical dependent variable of the original data was changed to a binomial category. For a particular choice of GLM, a measure of the model's predictive power can be useful for evaluating the practical importance of the predictors and for comparing competing GLMs, for example, models with different link functions or with different linear predictors.

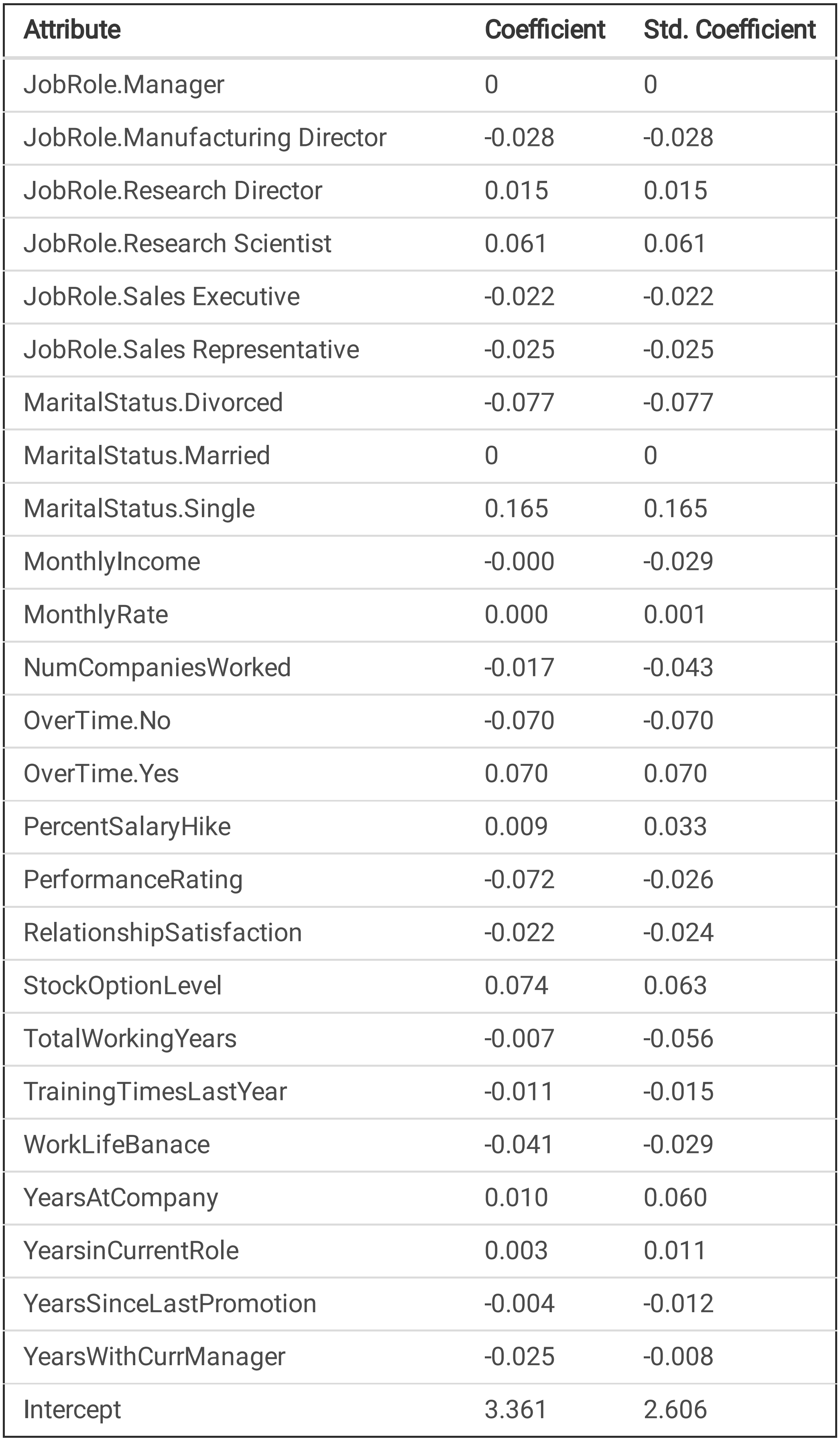
4.1 Linear regression model

In linear regression analysis, the model is expressed as a function. <Table 2 > shows the intercept, coefficient, and standard coefficient derived by regression analysis. It is the regression coefficient that explains how each explanatory variable affects the instep of the dependent variable. If a unit of measure with different explanatory variables is used, it is impossible to explain how an increase in one unit of explanatory variables affects the dependent variable. To solve this, we obtained the standard coefficient by estimating the regression model after standardizing the variables. Standard coefficients can be used to compare how each explanatory variable affects the dependent variable. In the < Table 2>, Age, Attrition.No, BusinessTravel.Travel_Frequently, DailyRate, Department.Sales, DistanceFromHome, EducationField.Life Sciences, EducationField.Other, EducationField.Technical Degree, Gender.Male, JobLevel, JobRole.Healthcare Representative, JobRole.Laboratory Technician, JobRole.Research Director, JobRole.Research Scientist, MaritalStatus.Single, MonthlyRate, OverTime.Yes, PercentSalaryHike, StockOptionLevel, YearsAtCompany, and YearsinCurrentRole are shown to increase JobSatisfaction.

<Table 2 > The results of the linear regression model



Attribute	Coefficient	Std. Coefficient
Age	0.003	0.025
Attrition.No	0.216	0.216
Attrition.Yes	-0.216	-0.216
BusinessTravel.Non-Travel	0	0
BusinessTravel.Travel_Frequently	0.073	0.073
BusinessTravel.Travel_Rarely	-0.044	-0.044
DailyRate	0.000	0.029
Department.Human Resources	0	0
Department.Research & Development	-0.097	-0.097
Department.Sales	0.033	0.033
DistanceFromHome	0.001	0.008
Education	-0.005	-0.005
EducationField.Human Resources	-0.025	-0.025
EducationField.Life Sciences	0.088	0.088
EducationField.Marketing	-0.081	-0.081
EducationField.Medical	-0.015	-0.015
EducationField.Other	0.033	0.033
EducationField.Technical Degree	0.003	0.003
EmployeeNumber	-0.000	-0.052
EnvironmentSatisfaction	-0.027	-0.027
Gender.Female	-0.046	-0.046
Gender.Male	0.046	0.046
HourlyRate	-0.004	-0.004
JobInvolvement	-0.053	-0.038
JobLevel	0.043	0.048
JobRole.Healthcare Representative	0.064	0.064
JobRole.Human Resources	-0.129	-0.129
JobRole.Laboratory Technician	0.022	0.022



Attribute	Coefficient	Std. Coefficient
JobRole.Manager	0	0
JobRole.Manufacturing Director	-0.028	-0.028
JobRole.Research Director	0.015	0.015
JobRole.Research Scientist	0.061	0.061
JobRole.Sales Executive	-0.022	-0.022
JobRole.Sales Representative	-0.025	-0.025
MaritalStatus.Divorced	-0.077	-0.077
MaritalStatus.Married	0	0
MaritalStatus.Single	0.165	0.165
MonthlyIncome	-0.000	-0.029
MonthlyRate	0.000	0.001
NumCompaniesWorked	-0.017	-0.043
OverTime.No	-0.070	-0.070
OverTime.Yes	0.070	0.070
PercentSalaryHike	0.009	0.033
PerformanceRating	-0.072	-0.026
RelationshipSatisfaction	-0.022	-0.024
StockOptionLevel	0.074	0.063
TotalWorkingYears	-0.007	-0.056
TrainingTimesLastYear	-0.011	-0.015
WorkLifeBanace	-0.041	-0.029
YearsAtCompany	0.010	0.060
YearsinCurrentRole	0.003	0.011
YearsSinceLastPromotion	-0.004	-0.012
YearsWithCurrManager	-0.025	-0.008
Intercept	3.361	2.606

Gaussian was used as the distribution function (family) when creating the model, and identity was used as the link function (link). Because the verification was performed as a cross-validation, it may appear differently for each subset. The linear regression model performance indicators in the < Table 3 > show root_mean_squared_error (1.116 +/- 0.034), absolute_error (0.956 +/- 0.031), relative_error (53.32% +/- 3.68%), squared_error (1.246 +/- 0.075), andcorrelation (0.064 +/- 0.066).

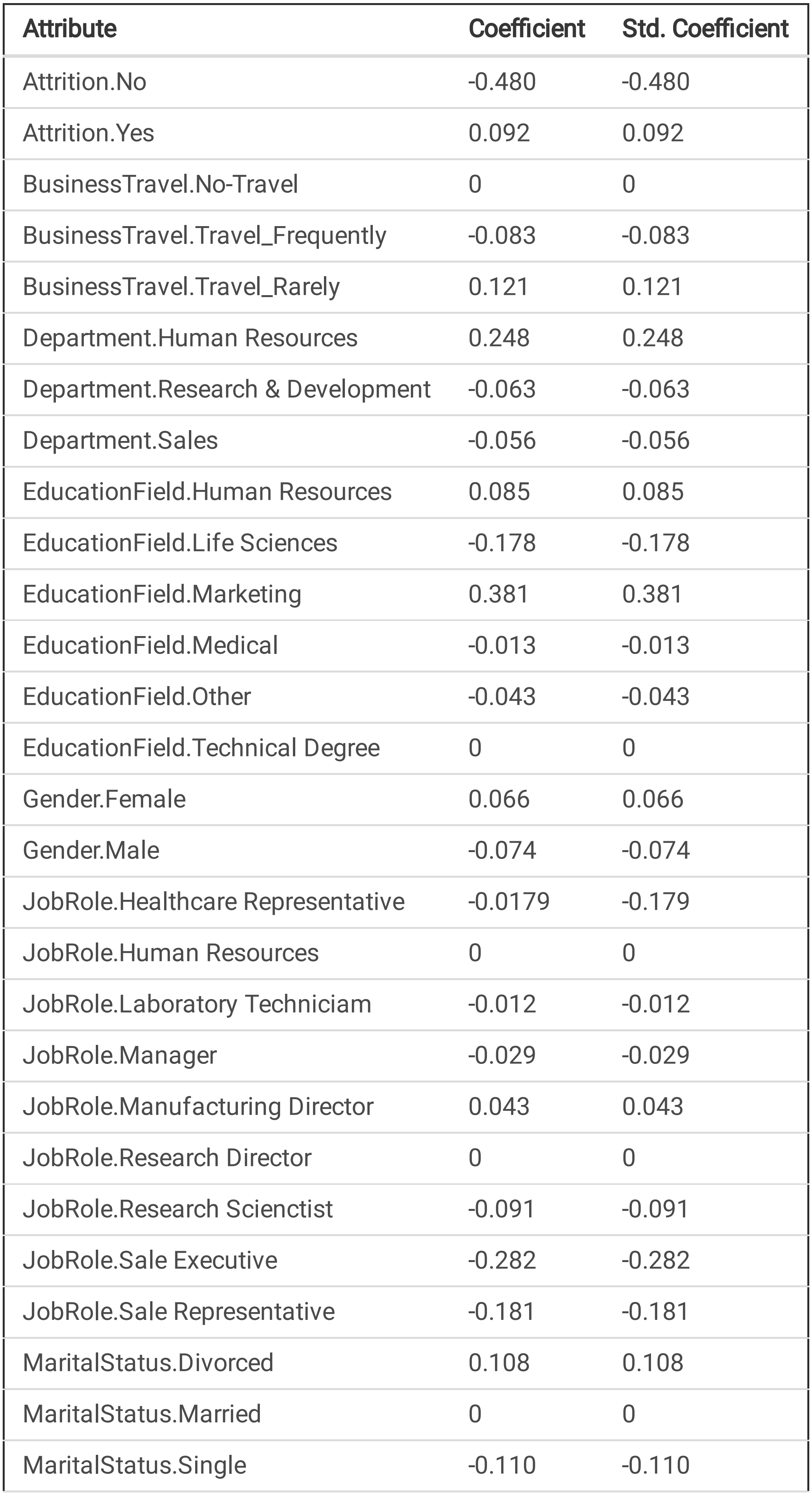
<Table 3 > The performance of the linear regression model

Performance indicator	Measurement value
root_mean_squared_error	1.116 +/- 0.034
absolute_error	0.956 +/- 0.031
relative_error	53.32% +/- 3.68%
squared_error	1.246 +/- 0.075
correlation	0.064 +/- 0.066

4.2 Binomial classification model

In the original data, JobSatisfaction is numerical data. For binomial classification, we create a property called JobSatisfaction 2 and create 'H' if the JobSatisfaction is greater than or equal to 3, and 'L' if it is less. In binomial classification, the model is expressed in the form of a function. <Table 4 > shows the intercept, coefficient, and standard coefficient derived by regression analysis. In the < Table 4>, Attrition.Yes, BusinessTravel.Travel_Rarely, Department.Human Resources, EducationField.Human Resources, EducationField.Marketing, Gender.Female, JobRole.Manufacturing Director, MaritalStatus.Divorced, NumCompaniesWorkedtrue, OverTime.No, StockOptionLevel.false, TotalWorkingYears.true, TrainingTimesLastYear.true, YearsAtCompany.false, YearsinCurrentRole.false, YearsSinceLastPromotion.false, and YearsWithCurrManager.true are shown to make more than 3 level of JobSatisfaction.

<Table 4 > The results of the binomial classification model



Attribute	Coefficient	Std. Coefficient
Attrition.No	-0.480	-0.480
Attrition.Yes	0.092	0.092
BusinessTravel.No-Travel	0	0
BusinessTravel.Travel_Frequently	-0.083	-0.083
BusinessTravel.Travel_Rarely	0.121	0.121
Department.Human Resources	0.248	0.248
Department.Research & Development	-0.063	-0.063
Department.Sales	-0.056	-0.056
EducationField.Human Resources	0.085	0.085
EducationField.Life Sciences	-0.178	-0.178
EducationField.Marketing	0.381	0.381
EducationField.Medical	-0.013	-0.013
EducationField.Other	-0.043	-0.043
EducationField.Technical Degree	0	0
Gender.Female	0.066	0.066
Gender.Male	-0.074	-0.074
JobRole.Healthcare Representative	-0.0179	-0.179
JobRole.Human Resources	0	0
JobRole.Laboratory Techniciam	-0.012	-0.012
JobRole.Manager	-0.029	-0.029
JobRole.Manufacturing Director	0.043	0.043
JobRole.Research Director	0	0
JobRole.Research Scienctist	-0.091	-0.091
JobRole.Sale Executive	-0.282	-0.282
JobRole.Sale Representative	-0.181	-0.181
MaritalStatus.Divorced	0.108	0.108
MaritalStatus.Married	0	0
MaritalStatus.Single	-0.110	-0.110

Attribute	Coefficient	Std. Coefficient
NumCompaniesWorkedfalse	-0.038	-0.038
NumCompaniesWorkedtrue	0.037	0.037
OverTime.No	0.114	0.114
OverTime.Yes	-0.121	-0.121
StockOptionLevel.false	0.026	0.026
StockOptionLevel.true	-0.026	-0.026
TotalWorkingYears.false	-1.346	-1.346
TotalWorkingYears.true	0.905	0.905
TrainingTimesLastYear.false	-0.305	-0.305
TrainingTimesLastYear.true	0.251	0.251
YearsAtCompany.false	0.188	0.188
YearsAtCompany.true	-0.254	-0.254
YearsinCurrentRole.false	0.016	0.016
YearsCurrentRole.true	-0.018	-0.018
YearsSinceLastPromotion.false	0.010	0.010
YearsSinceLastPromotion.true	-0.010	-0.010
YearsWithCurrManager.false	-0.066	-0.066
YearsWithCurrManager.true	0.072	0.072
Intercept	-0.919	-0.919

Binomial was used as the distribution function (family) when creating the model, and logit was used as the link function (link). The binomial classification model performance indicators in the < Table 5 > show accuracy (59.86% +/- 2.83%), AUC (0.519 +/- 0.057), precision (44.74% +/- 13.61%), recall (10.37% +/- 2.68%), and f_measure (16.62% +/-4.14%).

<Table 5 > The performance of the binomial classification model

Performance indicator	Measurement value
accuracy	59.86% +/- 2.83%
AUC	0.519 +/- 0.057
precision	44.74% +/- 13.61%
recall	10.37% +/- 2.68%
f_measure	16.62% +/-4.14%

5. Conclusions

This study identifies the factors determining job satisfaction. Job satisfaction has a huge impact on workplace productivity. A lot of studies have been reported about job satisfaction, but no one can say that we can create a universal human tool to predict job satisfaction. Job satisfaction is so complex and connected to so many elements that researchers tend to use fewer elements and ignore the effects of other factors.

The main purpose of this paper is to test the accuracy of models and develop a new model to predict job satisfaction. To recap, this study essentially had two primary goals. Firstly, this paper intends to understand the role of variables in job satisfaction prediction modeling better. Secondly, the study seeks to evaluate the predictive performance of the GLM including linear regression and binomial classification. Based on the findings reported above, a series of implications are drawn. Concerning the first goal, the findings of the study suggest that assessing the role of variables is complex and that their influences vary according to the types of GLM employed. The GLM highlights the explanatory power as most important to the analysis. Therefore, collectively no unanimous conclusions can be drawn about which explanatory variables are most critical to loan prediction for all the methods employed in totality. Yet, the findings of this study do shed some additional light on the employee's profile. The HR departments should be seeking to predict job satisfaction on GLM employed.

This study contributes to the literature regarding job satisfaction by providing a global model summarizing the job satisfaction of employees' demographics with machine learning. Machine learning techniques including linear regression and binomial classification along with feature importance analyses are employed to achieve the best results in terms of accuracy. The findings provide a comprehensive understanding of the job satisfaction determinants in workplace. Practically, this study provides insights for companies to manage job satisfaction. This paper attempts to come up with the best-performing model for predicting job satisfaction based on a limited set of features including employees' demographics. With this methodology, this study identified a pattern of employees' demographics that can predict job satisfaction. Moreover, this study can present specific task guidelines to the HR leaders who strive to increase job satisfaction, as they quantify the determinant factors that actually occur.

In the future, the machine learning model will make use of a larger training dataset, possibly more than a million different data points maintained in an electronic HR system. Although it would be a huge leap in terms of computational power and software sophistication, a system that will work on artificial intelligence might allow the HR leaders to decide the best-suited decision for the concerned employees as soon as possible.

Nevertheless, this study acknowledges an important limitation of this study. Economic modeling is used to explore the dataset and identify the associations between various factors and job satisfaction. However, social or psychological factors governing job satisfaction can be considered. Therefore, it will be important to conduct quantitative research to explore the rationale for job satisfaction.

Declarations

Funding: The authors did not receive support from any organization for the submitted work.

Competing interests: The authors have no relevant financial or non-financial interests to disclose.

Author contributions: All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Youngkeun Choi and Jae Won Choi. The first draft of the manuscript was written by Youngkeun Choi and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

References

1. Armstrong, M. (2006). A Handbook of Human resource Management Practice, Tenth Edition, Kogan Page Publishing, London,, p. 264
2. Ash A, & Shwartz M. (1999). R^2: a useful measure of model performance when predicting a dichotomous outcome. Statistics in Medicine, 18, 375–384.
3. Aziri, B. (2008). Menaxhimi i burimeve njerëzore, Satisfaksioni nga puna dhe motivimi i punëtorëve, Tringa Design, Gostivar, p. 46.
4. Bakry, U., Ayeldeen, H., Ayeldeen, G., & Shaker, O. (2016). Classification of Liver Fibrosis patients by multi-dimensional analysis and SVM classifier: an Egyptian case study. In: Proceedings of SAI Intelligent Systems Conference, pp. 1085–1095. Springer, Cham.
5. Boyd, D. & Crawford, K. (2012). Critical questions for big data. Provocations for a cultural, technological, and scholarly phenomenon. Information, Communication & Society, 15(5), 662–679.
6. Cherif, F. (2020). The role of human resource management practices and employee job satisfaction in predicting organizational commitment in Saudi Arabian banking sector. International Journal of Sociology and Social Policy, 40(7/8), 529–541.
7. Cox, D. R., & Wermuth, N. (1992). A comment on the coefficient of determination for binary responses. American Statistician, 46, 1–4.

8. Davis, K. & Nestrom, J. W. (1985). Human Behavior at work: Organizational Behavior, 7 edition, McGraw Hill, New York, p.109.
9. Durant, K. T., & Smith, M. D. (2006). Predicting the political sentiment of web log posts using supervised machine learning techniques coupled with feature selection. In: International Workshop on Knowledge Discovery on the Web, pp. 187–206. Springer, Berlin, Heidelberg.
10. Efron, B. (1978). Regression and anova with zero-one data: measures of residual variation. Journal of the American Statistical Association, 73, 113–121.
11. García, G. A., Gonzales-Miranda, D. R., Gallo, O. & Roman-Calderon, J. P. (2019). Employee involvement and job satisfaction: a tale of the millennial generation. Employee Relations, 41(3), 374–388.
12. George, J.M. and Jones, G.R. (2008). Understanding and Managing Organizational behavior, Fifth Edition, Pearson/Prentice Hall, New Yersey, p. 78
13. Haberman, S. J. (1982). Analysis of dispersion of multinomial responses. Journal of the American Statistical Association, 77, 568–580.
14. Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical tests. Journal of the American Medical Association, 247, 2543–2546.
15. Harrell, F. E. Jr, Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing error. Statistics in Medicine, 15, 361–387.
16. Hughes, C., Robert, L., Frady, K. & Arroyos, A. (2019). Prelims, Managing Technology and Middle- and Low-skilled Employees (The Changing Context of Managing People), Emerald Publishing Limited, pp. i-xvii.
17. Jacobs, A. (2009). Pathologies of Big Data. Communications of the ACM, 52(8), 36–44.
18. Jahan, S. (2014). Human Resources Information System (HRIS): A Theoretical Perspective. Journal of Human Resource and Sustainability Studies, 2(02), 33–39.
19. Kaliski, B.S. (2007). Encyclopedia of Business and Finance, Second edition, Thompson Gale, Detroit, p. 446.
20. Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. Big Data & Society, 1(1), 1–12.
21. Korn, E. L., & Simon, R. Measures of explained variation for survival data. Statistics in Medicine, 9, 487–503.
22. LaFayette, B., Curtis, W., Bedford, D. & Iyer, S. (2019). How Human Resource Management Changes in the Knowledge Economy. Knowledge Economies and Knowledge Work (Working Methods for Knowledge Management), Emerald Publishing Limited, pp. 161–169.
23. Li, Y. M., Lai, C. Y., & Kao, C. P. (2011). Building a qualitative recruitment system via SVM with MCDM approach. Appl. Intell. 35, 75–88.
24. Mauro, T. G. & Borges-Andrade, J. E. (2020), Human resource system as innovation for organisations, Innovation & Management Review, 17(2), 197–214.

25. Mathias, H. D., & Ragusa, V. R. (2016). Micro aerial vehicle path planning and flight with a multiobjective genetic algorithm. In Proceedings of SAI Intelligent Systems Conference, pp. 107–124. Springer, Cham.

26. Mayer-Schönberger, V. & Cukier, K. (2012), Big Data: A Revolution that will Transform how we Live, Work, and Think. New York, New York: Houghton Mifflin Harcourt.

27. Mittlbock M, & Schemper M. (1996). Explained variation for logistic regression. Statistics in Medicine, 15, 1987–1997.

28. Mullins, J.L. (2005). Management and organizational behavior, Seventh Edition, Pearson Education Limited, Essex, p. 700

29. Nunan, D. & Di Domenico, M. (2017). Big data: A normal accident waiting to happen? Journal of Business Ethics, 145(3), 481–491.

30. Spector, P.E. (1997). Job satisfaction: Application, assessment, causes and consequences,Thousand Oaks, CA,Sage Publications, Inc

31. Statt, D. (2004). The Routledge Dictionary of Business Management, Third edition, Routledge Publishing, Detroit, p. 78.

32. Theil, H. (1970). On the estimation of relationships involving qualitative variables. American Journal of Sociology, 76, 103–154.