Received 28 May 2020
Revised 1 June 2020
Accepted 31 July 2020

# A study of job involvement prediction using machine learning technique

Youngkeun Choi
*Sangmyung University, Jongno-gu, Republic of Korea, and*
Jae Won Choi
*Erik Jonsson School of Engineering and Computer Science, The University of Texas at Dallas, Richardson, Texas, USA*

## Abstract

**Purpose** – Job involvement can be linked with important work outcomes. One way for organizations to increase job involvement is to use machine learning technology to predict employees' job involvement, so that their leaders of human resource (HR) management can take proactive measures or plan succession for preservation. This paper aims to develop a reliable job involvement prediction model using machine learning technique.

**Design/methodology/approach** – This study used the data set, which is available at International Business Machines (IBM) Watson Analytics in IBM community and applied a generalized linear model (GLM) including linear regression and binomial classification. This study essentially had two primary approaches. First, this paper intends to understand the role of variables in job involvement prediction modeling better. Second, the study seeks to evaluate the predictive performance of GLM including linear regression and binomial classification.

**Findings** – In these results, first, employees' job involvement with a lot of individual factors can be predicted. Second, for each model, this model showed the outstanding predictive performance.

**Practical implications** – 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 HR management problems.

**Originality/value** – This paper is the first one to attempt to come up with the best-performing model for predicting job involvement based on a limited set of features including employees' demographics using machine learning technique.

**Keywords** Machine learning, Data science, Generalized linear model, Human resource management, Job involvement

**Paper type** Research paper

## 1. Introduction

Computer and internet-based technology development is a source of rapid increase in the amount and availability of data worldwide. Capture 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

International Journal of Organizational Analysis
Vol. 29 No. 3, 2021
pp. 788-800

1934-8835
DOI 10.1108/IJOA-05-2020-2222

(Mathias and Ragusa, 2016), political science (Durant and Smith, 2006) and many other areas. In response to advances in information technology, researchers studied a number of machine running approaches to improve human resources (HR) management outcomes (Li *et al.*, 2011).

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 and Di Domenico, 2017). Mayer-Schönberger and Cukier (2012) 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.

HR 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 s 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.

Job involvement can be linked with important work outcomes, such as being committed to their employer, heightened satisfaction from the job, increased work effort, reduced absenteeism, increased organizational citizenship (i.e. going above and beyond what is required at work) and reduced turnover intent and turnover (Chen and Chiu, 2009; Diefendorff *et al.*, 2002). One way for organizations to solve this problem is to use machine learning technology to predict employees' job involvement so that their leaders and HR can take proactive measures or plan succession for preservation. However, the machine learning technology historically used to solve this problem does not account for data noise in most HR information systems (HRIS). 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 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 involvement. 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 involvement, 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 used in the past (e.g. very few women are employed).

This paper describes the key machine learning algorithms used to address employees' job involvement issues. A new contribution to this paper is to explore the application of machine learning. 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 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

Early in the development of the concept of job involvement, there was a confusion over how to define it (Kanungo, 1979). Job involvement was first introduced by Lodahl and Kejner (1965), who conceptualized it as the degree to which a person is identified psychologically with his work or the importance of work in his total self-image. In addition, Lodahl and Kejner (1965) also saw job involvement as a result of how work performance affects a person's self-esteem. The conceptualization of dual dimensions of job involvement caused confusion on how to measure it in the beginning stages of research on the concept (Brown, 1996). Lawler and Hall (1970) contended that job involvement was the psychological identification with one's work and the degree to which the job situation is central to the person and his identity. The published works of Kanungo (1979) helped solidify the conceptualization of job involvement as a cognitive identification with the job. Kanungo argued that the lack of conceptual clarity derived from the fact that Lodahl and Kejner (1965) included two different concepts in their definition. Kanungo (1982) saw job involvement as only the cognitive identification with the job. The view of job involvement as job performance affecting the self-esteem of workers has fallen out of use (Brown, 1996).

Among contemporary researchers, the job involvement definition proposed by Kanungo (1982) is generally used. For example, Parasuraman (1982) conceptualized job involvement as the level of the person's ego involvement with their work. Elloy *et al.* (1992) saw job involvement as a generalized cognitive state of psychological identification with the job. Paullay *et al.* (1994) defined job involvement as when an employee is cognitively preoccupied with, engaged in and concerned with one's present job. All these and other current definitions of job involvement focus on the cognitive identification a person has with his/her job. Job involvement, therefore, is the degree of psychological identification a person has with the type of work that he or she is doing (Brown, 1996; Kanungo, 1982; Lawler and Hall, 1970). In other words, job involvement focuses on the degree of central interest the job plays in a person's life (i.e. the importance the person places on the job in his/her life) (Kanungo, 1979; Paullay *et al.*, 1994). As pointed out by DeCarufel and Schaan (1990), an individual with a high degree of job involvement would place the job at the center of his/her life's interests. The well-known phrase "I live, eat and breathe my job" would describe someone whose job involvement is very high.

Kanungo (1979) contended that the opposite of job involvement is job alienation. Job alienation is the feeling of being detached from the job and feeling that the job is unimportant in one's life. DeCarufel and Schaan (1990) pointed out that persons with low job involvement would place something other than their jobs (e.g. family and hobbies) at the center of their lives. For this study, the conceptualization of job involvement was the cognitive identification with the job proposed by Kanungo (1982). In the early research stages, there was criticism that job involvement overlapped with other workplace concepts,

such as work ethic, job satisfaction and organizational commitment (Brown, 1996). Theoretically, we assert that these are distinct concepts. Work ethic (formerly referred to as the Protestant work ethic) is the belief that work is important in the moral development of a person, and work will help make an individual better, as well as improve the character of the person. Thus, work ethic is a belief that work is important in the positive development of people (DeCarufel and Schaan, 1990) or the belief that work is important and that a person should work as hard as possible (Kanungo, 1982). Work ethic does not mean that a person psychologically identifies with his/her job but rather focuses on the general view that work is important for the development of human beings. It is, therefore, possible for a person to hold the belief that it is important to work but does not psychologically identify with the particular job held at the time.

Conversely, a person who identifies with his/her job (e.g. shapes his/her identity as being a correctional officer) may have a low level of work ethic and not put forth much effort toward his/her job; thus, it is also possible to have a low level of work ethic but a high level of job involvement. Besides being seen theoretically as distinct concepts, work ethic and job involvement have been empirically shown to be distinct concepts (Kanungo, 1982). Lawler and Hall (1970) were among the researchers to theorize that job involvement was different from job satisfaction. Locke (1976) defined job satisfaction as a pleasurable or positive emotional state resulting from the appraisal of one's job or job experiences. To Muchinsky (1987), job satisfaction was an emotional, affective response resulting from the extent a person derives pleasure from his/her job. Spector (1996) contended that job satisfaction was simply the extent to which people like their jobs. Therefore, job satisfaction is the emotional satisfaction from the job, while job involvement is the cognitive identification with the job (Kanungo, 1982). As Brooke *et al.* (1988) pointed out, job satisfaction is the emotional state of liking one's job, while job involvement is the cognitive belief state of psychological identification with one's job. It is, therefore, possible for an individual to identify with the job, regardless if he or she derives pleasure from the job (Kanungo, 1982). For example, a person could build his/her core self-view around being a correctional officer, but do not gain satisfaction from being assigned to work in a security tower.

Likewise, a person could find satisfaction from the job because of pay but not cognitively identify being a correctional officer. This individual might be willing to switch occupations if another job met his/her emotional needs, such as taking a job at a new manufacturing plant that paid a higher salary. Finally, a person could identify with the job of a correctional officer and gain satisfaction from the job. Hence, theoretically, job satisfaction and job involvement are separate concepts. In addition, research has empirically demonstrated that they are separate concepts. Using factor analysis, Lawler and Hall (1970) found that work ethic, job satisfaction and job involvement were empirically distinct concepts. Other studies have also found that job satisfaction and job involvement are distinct concepts (Brooke *et al.*, 1988). Organizational commitment and job involvement are also unique concepts. Organizational commitment is the bond between the employee and the agency (Mowday *et al.*, 1982). It is a bond with the entire employing organization and not with the job itself or a particular part of the organization (Lambert *et al.*, 1999). Organizational commitment is generally defined as having the core elements of loyalty to the organization, identification with the organization and involvement in the organization (Mowday *et al.*, 1982). Thus, organizational commitment is the identification with the overall employing organization, and job involvement is the identification with the job. Organizational commitment, therefore, focuses on bond at the organizational level and job involvement focuses on the attachment at the job level (Brown, 1996; Kanungo, 1982). Blau (1987) empirically demonstrated the discriminant validity of the concepts of job involvement and

organizational commitment. Additionally, using factor analytic procedures, Brooke *et al.* (1988) empirically demonstrated that job involvement, job satisfaction and organizational commitment were distinct concepts from one another. Moreover, they observed that different aspects of the work environment affected each of the concepts at differing degrees, again empirically indicating that they were separate concepts.

## 3. Methodology

### *3.1 Data set*

The data set used in this paper was related to job involvement and is available at international business machines (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 are still with my company or whether they have gone to work somewhere else. In addition, Watson Analytics wants to understand how this relates to workforce attrition. For each of 10 years, it show 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.

### *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 a 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 and 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 $\tilde{Y}$ and the outcomes $Y$ are concordant, the observation with the larger $Y$ also having the larger $\tilde{Y}$. For a binary response, $c$ is related to a widely used measure of diagnostic discrimination, the area under a receiver operating characteristic 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 $c$ uses 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.

| Variables | Measurement |
|---|---|
| Age | Integer |
| Attrition | Binomial (true or false) |
| BusinessTravel | Polynomial (Travel_Rarely 71%, Travel_Frequently 19% and other 10%) |
| DailyRate | Integer |
| Department | Polynomial (research and development 65%, sales 30% and other 4%) |
| DistanceFromHome | Integer |
| Education | Integer (1 "below college," 2 "college," 3 "bachelor," 4 "master" and 5 "doctor") |
| EducationField | Polynomial (life sciences 41%, medical 32% and other 27%) |
| EmployeeCount | Integer |
| EmployeeNumber | Integer |
| EnvironmentSatisfaction | Integer |
| Gender | Binomial (male 60% and female 40%) |
| HourlyRate | Integer |
| JobLevel | Integer |
| JobRole | Polynomial (Sales Executive 22%, Research Scientist 20% and other 58%) |
| JobSatisfaction | Integer (1 "low," 2 "medium," 3 "high," and 4 "very high") |
| MartialStatus | Polynomial (married 46%, single 32% and 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" and 4 "outstanding") |
| RelationshipSatisfaction | Integer (1 "low," 2 "medium," 3 "high" and 4 "very high") |
| StandardHours | Integer |
| StockOptionLevel | Integer |
| TotalWorkingYears | Integer |
| TrainingTimesLastYear | Integer |
| WorkLifeBalance | Integer (1 "bad," 2 "good," 3 "better" and 4 "best") |
| YearsAtCompany | Integer |
| YearsinCurrentRole | Integer |
| YearsSinceLastPromotion | Integer |
| YearsWithCurrManager | Integer |
| JobInvolvement | Integer (1 "low," 2 "medium," 3 "high" and 4 "very high") |

**Table 1.**
The measurements of variables

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 machine learning 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:

$$R^2 = 1 - \left[\sum_{i=1}^{n} \left(y_i - \widehat{\Pi}_i\right)^2\right] / \left[\sum_{i=1}^{n} (y_i - \bar{y})^2\right]$$

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 and Wermuth, 1992). Others have argued, however, 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 and Shwartz, 1999) and that $R^2$ captures information (Ash and Shwartz, 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 and 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% 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 Pre-processing and data mining models*

In this study, we want to analyze the factors in the effect on job development. 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, namely, numerical prediction, binomial classification. Therefore, the numerical dependent variable of the original data was changed to a binomial category. Statistical and data mining techniques have been used 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 data set and the test data set are separated for machine learning, the test data set must satisfy the following requirements. First, the training data set and the test data set must be created in the same format. Second, the test data set should not be included in the training data set. Third, the training data set and the test data set 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 data set 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 data set into a training data set and a test data set 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.

| Attribute | Coefficient | Std. coefficient |
|---|---|---|
| Age | 0.003 | 0.003 |
| Attrition.No | 0.139 | 0.139 |
| Attrition.Yes | −0.139 | −0.139 |
| BusinessTravel.Non-Travel | −0.143 | −0.143 |
| BusinessTravel.Travel_Frequently | 0.032 | 0.032 |
| BusinessTravel.Travel_Rarely | 0 | 0 |
| DailyRate | 0.000 | 0.026 |
| Department.Human Resources | 0.129 | 0.129 |
| Department.Research and Development | −0.023 | −0.023 |
| Department.Sales | −0.111 | −0.111 |
| DistanceFromHome | 0.002 | 0.014 |
| Education | 0.025 | 0.026 |
| EducationField Life Sciences | 0 | 0 |
| EducationField.Human Resources | 0 | 0 |
| EducationField.Marketing | −0.018 | −0.018 |
| EducationField.Medical | 0.018 | 0.018 |
| EducationField.Other | −0.047 | −0.047 |
| EducationField.Technical Degree | 0.011 | 0.011 |
| EmployeeNumber | −0.000 | −0.003 |
| EnvironmentSatisfaction | −0.014 | −0.015 |
| Gender.Female | −0.020 | −0.020 |
| Gender.Male | 0.020 | 0.020 |
| HourlyRate | 0.001 | 0.024 |
| JobLevel | 0.052 | 0.057 |
| JobRole.Healthcare Representative | −0.015 | −0.015 |
| JobRole.Human Resources | −0.225 | −0.225 |
| JobRole.Laboratory Technician | −0.093 | −0.093 |
| JobRole.Manager | 0.348 | 0.348 |
| JobRole.Manufacturing Director | −0.077 | −0.077 |
| JobRole.Research Director | 0.235 | 0.235 |
| JobRole.Research Scientist | −0.017 | −0.017 |
| JobRole.Sales Executive | 0.074 | 0.074 |
| JobRole.Sales Representative | 0.010 | 0.010 |
| JobSatisfaction | −0.022 | −0.024 |
| MaritalStatus.Divorced | 0.007 | 0.007 |
| MaritalStatus.Married | 0 | 0 |
| MaritalStatus.Single | −0.047 | −0.047 |
| MonthlyIncome | 0 | −0.157 |
| MonthlyRate | 0 | −0.008 |
| NumCompaniesWorked | 0.005 | 0.014 |
| OverTime.No | −0.026 | −0.026 |
| OverTime.Yes | 0.026 | 0.026 |
| PercentSalaryHike | 0.002 | 0.007 |
| PerformanceRating | −0.077 | −0.028 |
| RelationshipSatisfaction | 0.018 | 0.019 |
| StockOptionLevel | −0.021 | −0.018 |
| TotalWorkingYears | −0.003 | −0.025 |
| TrainingTimesLastYear | −0.011 | −0.015 |
| WorkLifeBalance | −0.019 | −0.013 |
| YearsAtCompany | −0.011 | −0.070 |
| YearsinCurrentRole | 0.006 | 0.023 |
| YearsSinceLastPromotion | −0.003 | −0.009 |
| YearsWithCurrManager | 0.017 | 0.062 |
| Intercept | 2.884 | 2.689 |

**Table 2.**
The results of linear regression model

## 4. Results

### *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, the standard coefficient is obtained 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 Table 2, Age, Attrition.No, BusinessTravel.Travel_Frequently, DailyRate, Department.HR, DistanceFromHome, Education, EducationField.Medical, EducationField.Technical Degree, Gender.Male, HourlyRate, JobLevel, JobRole.Manager, JobRole.Research Director, JobRole.Sales Executive, JobRole.Sales Representative, MaritalStatus.Divorced, NumCompaniesWorked, OverTime.Yes, PercentSalaryHike, RelationshipSatisfaction, YearsinCurrentRole and YearsWithCurrManager are shown to increase JobInvolvement.

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 are as follows (Table 3).

### *4.2 Binomial classification model*

In the original data, JobInvolvement is numerical data. For binomial classification, we create a property called JobInvolvement2 and create "H" if the JobInvolvement 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.No-Travel, EducationField. HR, Gender.Female, JobRole.Healthcare Representative, JobRole.Laboratory Techniciam, JobRole.Manufacturing Director, MaritalStatus.Single, NumCompaniesWorkedfalse, OverTime.No, StockOptionLevel.true, TotalWorkingYears.true, TrainingTimesLastYear. true, YearsAtCompany.true, YearsCurrentRole.true, YearsSinceLastPromotion.true, YearsWithCurrManager.false are shown to make more than 3 level of JobInvolvement.

Binomial was used as the distribution function (family) when creating the model, and logit was used as the link function (link). Table 5 shows the overall performance indicators.

## 5. Conclusions

### *5.1 Discussion*

This study identifies the factors determining job involvement. Job involvement has a huge impact on workplace productivity. A lot of studies have been reported about job involvement, but no one can say that we can create a universal human tool to predict job

**Table 3.**
The performance of linear regression model

| Performance indicator | Measurement value |
|---|---|
| root_mean_squared_error | 0.716% +/− 0.037% |
| absolute_error | 0.560% +/− 0.038% |
| relative_error | 27.03% +/− 2.16% |
| squared_error | 0.514% +/− 0.053% |
| correlation | 0.085% +/− 0.076% |

| Attribute | Coefficient | Std. coefficient |
|---|---|---|
| Attrition.No | −0.359 | −0.359 |
| Attrition.Yes | 0.238 | 0.238 |
| BusinessTravel.No-Travel | 0.297 | 0.297 |
| BusinessTravel.Travel_Frequently | −0.109 | −0.109 |
| BusinessTravel.Travel_Rarely | 0 | 0 |
| Department.Human Resources | −0.085 | −0.085 |
| Department.Research and Development | 0 | 0 |
| Department.Sales | 0 | 0 |
| EducationField.Human Resources | 0.288 | 0.288 |
| EducationField.Life Sciences | −0.086 | −0.086 |
| EducationField.Marketing | 0 | 0 |
| EducationField.Medical | −0.057 | −0.057 |
| EducationField.Other | 0.088 | 0.088 |
| EducationField.Technical Degree | −0.128 | −0.128 |
| Gender.Female | 0.052 | 0.052 |
| Gender.Male | −0.054 | −0.054 |
| JobRole.Healthcare Representative | 0.202 | 0.202 |
| JobRole.Human Resources | 0 | 0 |
| JobRole.Laboratory Techniciam | 0.039 | 0.039 |
| JobRole.Manager | −0.169 | −0.169 |
| JobRole.Manufacturing Director | 0.262 | 0.262 |
| JobRole.Research Director | −0.030 | −0.030 |
| JobRole.Research Scienctist | −0.094 | −0.094 |
| JobRole.Sale Executive | 0 | 0 |
| JobRole.Sale Representative | −0.114 | −0.114 |
| MaritalStatus.Divorced | −0.144 | −0.144 |
| MaritalStatus.Married | 0 | 0 |
| MaritalStatus.Single | 0.152 | 0.152 |
| NumCompaniesWorkedfalse | 0.020 | 0.020 |
| NumCompaniesWorkedtrue | −0.021 | −0.021 |
| OverTime.No | 0.056 | 0.056 |
| OverTime.Yes | −0.055 | −0.055 |
| StockOptionLevel.false | −0.053 | −0.053 |
| StockOptionLevel.true | 0.055 | 0.055 |
| TotalWorkingYears.false | −0.905 | −0.905 |
| TotalWorkingYears.true | 0.790 | 0.790 |
| TrainingTimesLastYear.false | −0.219 | −0.219 |
| TrainingTimesLastYear.true | 0.211 | 0.211 |
| YearsAtCompany.false | −0.176 | −0.176 |
| YearsAtCompany.true | 0.178 | 0.178 |
| YearsinCurrentRole.false | −0.024 | −0.024 |
| YearsCurrentRole.true | 0.027 | 0.027 |
| YearsSinceLastPromotion.false | −0.014 | −0.014 |
| YearsSinceLastPromotion.true | 0.014 | 0.014 |
| YearsWithCurrManager.false | 0.209 | 0.209 |
| YearsWithCurrManager.true | −0.267 | −0.267 |
| Intercept | −1.519 | −1.519 |

**Table 4.**
The results of binomial classification model

involvement. Job involvement 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 involvement. To recap, this study essentially had two primary goals.

First, this paper intends to understand the role of variables in job involvement prediction modeling better. Second, 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 used. 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 used 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 involvement on GLM used.

### *5.2 Research contributions and practical implications*

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

### *5.3 Limitations and future research directions*

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

In the future, the machine learning model will make use of a larger training data set, 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.

**Table 5.**
The performance of linear regression model

| Performance indicator | Measurement value |
|---|---|
| accuracy | 67.69% +/− 0.98% |
| AUC | 0.524% +/− 0.063% |
| precision | 26.29% +/− 19.03% |
| recall | 3.50% +/− 2.95% |
| f_measure | 6.31% |

## References

Ash, A. and Shwartz, M. (1999), "$R^2$: a useful measure of model performance when predicting a dichotomous outcome", *Statistics in Medicine*, Vol. 18 No. 4, pp. 375-384.

Bakry, U., Ayeldeen, H., Ayeldeen, G. and 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*.

Blau, G. (1987), "Using a person-environment fit model to predict job involvement and organizational commitment", *Journal of Vocational Behavior*, Vol. 30 No. 3, pp. 240-257.

Boyd, D. and Crawford, K. (2012), "Critical questions for big data", *Provocations for a Cultural, Technological, and Scholarly Phenomenon. Information, Communication and Society*, Vol. 15 No. 5, pp. 662-679.

Brooke, P., Russell, D. and Price, J. (1988), "Discriminant validation of measures of job satisfaction, job involvement, and organizational commitment", *Journal of Applied Psychology*, Vol. 73 No. 2, pp. 139-145.

Brown, S. (1996), "A meta-analysis and review of organizational research in job involvement", *Psychological Bulletin*, Vol. 120 No. 2, pp. 235-255.

Chen, C.C. and Chiu, S.F. (2009), "The mediating role of job involvement in the relationship between job characteristics and organizational citizenship behavior", *The Journal of Social Psychology*, Vol. 149 No. 4, pp. 474-494.

Cox, D.R. and Wermuth, N. (1992), "A comment on the coefficient of determination for binary responses", *American Statistician*, Vol. 46, pp. 1-4.

DeCarufel, A. and Schaan, J.-L. (1990), "The impact of compressed work weeks on police job involvement", *Canadian Police College*, Vol. 14, pp. 81-97.

Diefendorff, J., Brown, D., Kamin, A. and Lord, R. (2002), "Examining the roles of job involvement and work centrality in predicting organizational citizenship behaviors and job performance", *Journal of Organizational Behavior*, Vol. 23 No. 1, pp. 93-108.

Durant, K.T. and 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*.

Efron, B. (1978), "Regression and Anova with zero-one data: measures of residual variation", *Journal of the American Statistical Association*, Vol. 73 No. 361, pp. 113-121.

Elloy, D., Everett, J. and Flynn, W. (1992), "An examination of the correlates of job involvement", *Group and Organization Studies*, Vol. 16 No. 2, pp. 160-177.

Haberman, S.J. (1982), "Analysis of dispersion of multinomial responses", *Journal of the American Statistical Association*, Vol. 77 No. 379, pp. 568-580.

Harrell, F.E., Jr, Lee, K.L. and Mark, D.B. (1996), "Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing error", *Statistics in Medicine*, Vol. 15 No. 4, pp. 361-387.

Harrell, F.E., Califf, R.M., Pryor, D.B., Lee, K.L. and Rosati, R.A. (1982), "Evaluating the yield of medical tests", *Jama: The Journal of the American Medical Association*, Vol. 247 No. 18, pp. 2543-2546.

Hilden, J. (1991), "The area under the ROC curve and its competitors", *Medical Decision Making*, Vol. 11 No. 2, pp. 95-101.

Jacobs, A. (2009), "Pathologies of big data", *Communications of the ACM*, Vol. 52 No. 8, pp. 36-44.

Jahan, S. (2014), "Human resources information system (HRIS): a theoretical perspective", *Journal of Human Resource and Sustainability Studies*, Vol. 2 No. 2, pp. 33-39.

Kanungo, R. (1979), "The concepts of alienation and involvement revisited", *Psychological Bulletin*, Vol. 86 No. 1, pp. 119-138.

Kanungo, R. (1982), *Work Alienation: An Integrative Approach*, Praeger, New York, NY.

Kitchin, R. (2014), "Big data, new epistemologies and paradigm shifts", *Big Data and Society*, Vol. 1 No. 1, pp. 1-12.

Korn, E.L. and Simon, R. (1990), "Measures of explained variation for survival data", *Statistics in Medicine*, Vol. 9 No. 5, pp. 487-503.

Lambert, E., Barton, S. and Hogan, N. (1999), "The missing link between job satisfaction and correctional staff behavior: the issue of organizational commitment", *American Journal of Criminal Justice*, Vol. 24 No. 1, pp. 95-116.

Lawler, E. and Hall, D. (1970), "Relationship of job characteristics to job involvement, satisfaction, and intrinsic motivation", *Journal of Applied Psychology*, Vol. 54 No. 4, pp. 305-312.

Li, Y.M., Lai, C.Y. and Kao, C.P. (2011), "Building a qualitative recruitment system via SVM with MCDM approach", *Applied Intelligence*, Vol. 35 No. 1, pp. 75-88.

Locke, E. (1976), "The nature and causes of job satisfaction", In M. Dunnette (Ed.), *Handbook of Industrial and Organizational Psychology*, (pp. 1297-1349). Rand-McNally, Chicago.

Lodahl, T. and Kejner, M. (1965), "The definition and measurement of job involvement", *Journal of Applied Psychology*, Vol. 49 No. 1, pp. 24-33.

Mathias, H.D. and 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.

Mayer-Schönberger, V. and Cukier, K. (2012), "Big data: a revolution that will transform how we live", *Work, and Think*, Houghton Mifflin Harcourt, New York, NY.

Mittlbock, M. and Schemper, M. (1996), "Explained variation for logistic regression", *Statistics in Medicine*, Vol. 15 No. 19, pp. 1987-1997.

Mowday, R., Porter, L. and Steers, R. (1982), *Employee-Organization Linkages: The Psychology of Commitment, Absenteeism, and Turnover*, Academic Press, New York, NY.

Muchinsky, P. (1987), *Psychology Applied to Work: An Introduction to Industrial and Organizational Psychology*, (2th ed.). Dorsey Press, Chicago.

Nunan, D. and Di Domenico, M. (2017), "Big data: a normal accident waiting to happen?", *Journal of Business Ethics*, Vol. 145 No. 3, pp. 481-491.

Parasuraman, S. (1982), "Predicting turnover intentions and turnover behavior: a multivariate analysis", *Journal of Vocational Behavior*, Vol. 21 No. 1, pp. 111-121.

Paullay, I., Alliger, G. and Stone-Romero, E. (1994), "Construct validation of two instruments designed to measure job involvement and work centrality", *Journal of Applied Psychology*, Vol. 79 No. 2, pp. 224-228.

Spector, P. (1996), *Industrial and Organizational Psychology: Research and Practice*, John Wiley, New York, NY.

Theil, H. (1970), "On the estimation of relationships involving qualitative variables", *American Journal of Sociology*, Vol. 76 No. 1, pp. 103-154.

**Corresponding author**
Youngkeun Choi can be contacted at: penking1@smu.ac.kr