OPEN ACCESS

International
Journal
Of Modern Engineering Research (IJMER)

A comparison of machine learning approaches for Employee Satisfaction Prediction

Haoyue Gao[1], Miao He[2], Guoxiang Hou[1]

[1]*School of Naval Architecture and Ocean Engineering, Huazhong University of Science and Technology, Wuhan, 430074, Hubei, China*

[2]*School of MechanicalScience and Engineering, Huazhong University of Science and Technology, Wuhan, 430074, Hubei, China*

ABSTRACT: *In this paper, we study three machine learning algorithms to predict the employee satisfaction based on a datas etextracted from the Kesci website. We also analyze various encoding methods which are used totrans form categorical data tonumerical data. Weuse linear regression, random forest regression and extreme gradient boosting (XGB) regression to train a model to predict the employe esatisfaction. The models are evaluated by using the state-of-art metricsfortheregression problem. The resultsare also numerically and visually presented.*

KEY WARDS:*Employee Satisfaction Prediction, Linear Regression, Random Forrest, XGB.*

Date of Submission: 08-11-2020 Date of acceptance: 23-11-2020

I. INTRODUCTION

The job turnover is an important topic the HR management tries to understand. It has been studied thoroughly by industry andmany researchers. There are two types of job turnover: voluntary and involuntary. The former is defined as the active action by the employee to leave the organization due to internal or external factors. The later involves the action carried out by the organization to end the contract with the employee [1]. Among all the factors that leads to job turnover, job satisfaction is by all means an undeniable aspect. It is suggested that job motivation and job satisfaction are among the factors that affect the job turnover [2]. It is also suggested that job satisfaction has a mediated effect on turnover intention and turnover intention has been identified as the most immediate cognitive antecedent to turnover [3]. With only satisfied and motivated employees, a company is able to achieve global success by producing world-class products [4].

Human resource management uses different methods to deal with prediction problem. Traditionally, structure equation model has been constructed and the questionnaire survey is used to empirically evaluate the model [4]. With the evolving of the machine learning theory and fast-growing digitally generated data, data mining techniques such as linear regression, decision tree and neural network has been implemented to solve the prediction problem such as bankruptcy forecast and services performance [5,6]. There are many machine learning models to be chosen from to fulfill the purpose of prediction. In order to mitigate the effect of contaminated data which is normally the case of the collection from HR departments, XGBoost algorithm is explored [7]. This conclusion is also backed up by Zhao [8] who investigates ten different machine learning methods by various data sources and types.Another concern for the dataset is that it does not satisfy the linearity assumption [9]. As a result, a sequential manifold learning model is proposed to solve the problem of the existing dimensionality reduction methods.Additionally, various studies from the area of Psychology and Management has been focused on the elements that impacts the job satisfaction.

In this paper, we analyze three common machine learning methods, linear regression, random forest and extreme gradient boost, and try to find out experimentally which is a better solution to deal with a specific regression problem that is to predict the employee satisfaction rate.

The paper is organized as the following. In section 2, we first sketch the model to be studied and then in Section3 we present the data analysis pipeline including exploratory data analysis, data pre-processing and evaluation metrics. Lastly, we present the results in Section 4.

II. METHODS AND CALCULATION

In this section, we present the machine learning models that are used to predict the job satisfaction.

2.1 Linear Regression

According to the publication in [10], a general linear regression model can be expressed as:

$$\hat{y} = w_0 + w_1 x_{i1} + w_2 x_{i2} + \cdots + w_n x_{in}, \tag{1}$$

where w_i represents the coefficient of the i-th attribute and x_{ji} represents the j-th instance value of the i-th attribute. By minimizing the error between the predicted output and the actual output as in the equation (2), the coefficients of the linear regression model can be determined by using the least-square estimator.

$$\text{error} = \sum_{i=0}^{m} (y - \hat{y})^2 \tag{2}$$

2.2 Random Forest

According to the publication in [11], the random forest is a decision tree based ensemble learning algorithm that uses bootstrapped dataset and use aggregate to make a decision.

2.3 *Extreme Gradient Boosting (XGB)*

According to the publication in [12], the extreme gradient boosting algorithm propose a scalable end-to-end tree boosting system for approximate tree learning. The object of this algorithm is to build a tree by optimizing the following function:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t), \tag{3}$$

where l is a differentiable convex loss function, $\hat{y}_i^{(t)}$ is the prediction of the i-th instance at t-th iteration, f_t represents the output of the regression tree and Ω is the regularization term. The equation (3) can be further approximated as:

$$\tilde{\mathcal{L}}^{(t)} = \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t), \tag{4}$$

III. DATA

In this section, we firstly present the data analysis of the dataset. Then we illustrate the data preprocessing step to prepare the all numerical dataset for the model to learn. Next, we introduce the evaluation metric to analyze the performance of each model.

3.1 Exploratory Data Analysis

The early data analysis (EDA) is used to inspect the dataset to provide an overview or spot any outlier by plotting or calculating statistic measures. It is an essential first step to understand the attributes and instances of the dataset for choosing the appropriate prediction model. The dataset is retrieved from the Kesci website [13]. The features of the dataset are listed in the table 1. The dataset has 12k instances and each instance consists of 10 features, among which 2 features are of the float type, 2 features are of the boolean type, 3 features are of the integer type and 3 features are of the categorical type. The names of each attribute are employee id, the score of the last evaluation, the number of the project completed by the employee, average monthly working hours, the length of service for the company, accidents at work, the level of package, promotion in last 5 years, division of the employee and the level of salaryrespectively. There are no missing or null value in the dataset and thus there is no need for the missing value inputation.In table 2, we present the typical value, i.e. central tendency the and its corresponding uncertainty, i.e. spread for the numerical features. As for the categorical features, we count the number of distinguished instances and found out that there are in total of 9 type of divisions, 3 level of salary and 5 level of packages.

***Table I*.Dataset Features**

#	Feature	Data Type
1	id	INT
2	last evaluation	FLOAT
3	number project	INT
4	average monthly hours	FLOAT
5	time spend company	INT
6	Work accident	BOOLEAN
7	package	CATEGORICAL
8	Promotion in last 5 years	BOOLEAN
9	division	CATEGORICAL
10	salary	CATEGORICAL

***Table II*.Analysis for integer and float type attributes**

Feature number	2	3	4	5
Central Tendency	0.72	3.79	202.2	3.5
Uncertainty	0.029	1.51	2488.8	2.14

Notice that the average-monthly-hours attribute in table 2illustrates a significant shift of central tendency and uncertainty comparing to other attributes. In order to provide a consistent distribution for the dataset, we transform this attribute to be within range [0, 1] by using the min-max normalization as shown in (5). After the transformation, we are able to shrink down the central tendency and uncertainty for the average-monthly-hours attribute to be 0.49 and 0.052 respectively.

$$\text{Min-Max Normalization} = \frac{X_{ij} - \min A_j}{\max A_j - \min A_j}, \tag{5}$$

where X denotes the dataset, i, j denotes the row and column index of the dataset X, and A denotes the feature of the dataset X.

3.2 Data preprocessing

The three models we are building to solve the prediction problem requires the input to be of numerical type. Therefore, we use various encoding methods to transfer the categorical data from texture form to numerical form. Given the 3 categorical features and 7 numerical features in the dataset, we drop the id feature and apply 5 different encoding methods to fulfill the requirement. Firstly, let's briefly introduce each encoding mothod in the following context:

3.2.1Label Encoding

It is a simple method to convert the cardinality (value) in a feature into a number.

3.2.2 One-Hot Encoding

For each categorical feature, the one-hot encoding firstly counts the number of cardinalities in that feature. Then, the individual value is assigned as bit 1 and the rest positions are embedded with 0. Given the 3 categorical features which produces 27 different cardinalities in the dataset, the number of columns of the features after the one-hot encoding will increase from 9 to 33, which is a significant change.

3.2.3 Target Encoding

Target encoding is the process of replacing a categorical value with the mean of the target variable. The advantage of this transformation is that it only takes one column of space. However, it is sensitive to the target variable and it is likely to overfit the model [14].

3.2.4 Weight-Of-Evidence (WOE) Encoding

Commonly used in in the credit and financial industry [15], Weight-of-Evidence encoding is a measure of separation of good and bad events. The definition can be found in (6).

$$WOE = \ln\left(\frac{\text{Distribution of Goods}}{\text{Distribution of Bads}}\right). \tag{6}$$

3.3Evaluation

We use several metrics to evaluate the performance of the model. Each metric is briefly introduced in the following context.

3.2.4 MeanSquared Error (MSE)

$$\mathrm{MSE} = \frac{1}{n}\sum (y - \hat{y})^2. \tag{7}$$

3.2.4 Root Mean Square Error (RMSE)

$$\mathrm{RMSE} = \sqrt{\frac{\sum (\hat{y} - y)^2}{n}}. \tag{8}$$

3.2.4 Mean Absolute Error (MAE)

$$\mathrm{MAE} = \frac{1}{n}\sum |y - \hat{y}|. \tag{9}$$

3.2.4 Coefficientof Determination R^2

$$\text{R-Square} = \frac{\sum (\hat{y} - \bar{y})^2}{\sum (y - \bar{y})^2}. \tag{10}$$

where y denotes the output value of the dataset, $\hat{y}$ denotes the predicted output value of the dataset, $\bar{y}$ denotes the expected value of y.

IV. RESULTS

We use various encoding techniques such as One-Hot encoding, Label encoding, Target encoding and Weight-of-Evidence encoding (WoE), which transform the categorical data to the numerical data. It is important to realize the impact each encoding method has on the training process since the encoding transformation will likely causes the sparsity of data which slows down the training process. The impact each encoding method exerted on the performance of the individual model is illustrated by the execution time fixing all the other parameters. The results can be found in the table 3.

***Table III*.Analysis for integer and float type attributes**

Method	LR	RF	XGB
One-Hot	1.98s	12.5s	17.2s
Label	1.70s	8.20s	7.53s
Target	1.79s	9.22s	9.89s
Weight-Of-Evidence	1.60s	10.3s	11.6s

The execution time of training three models by using four encoding methods are plotted in the figure 1. We notice that the One-Hot encoding consumes the longest execution time for all three algorithms because the One-Hot encoding dilutes the volume of the data attributes proportionally to the amount of the distinguished features of that attribute. As a result, more computation resource is needed during the training process. The other three encoding methods, however, does not take up additional space in the dataset and therefore, significantly consumes less time. In accordance with the complexity of the remaining three encoding methods, their corresponding length of execution time ranks as follows, the Label encoding, the Target encoding and the WoE encoding.

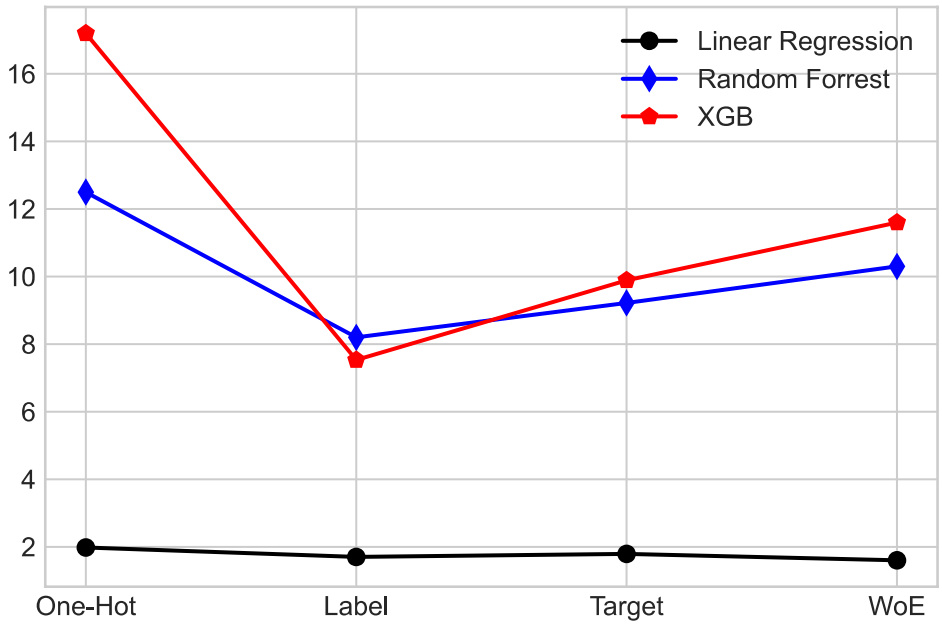

Figure 1: Execution time for 3 models

In addition to the execution time, the encoding methods also affects the performance of the model in terms of the prediction accuracy. The results for the performance of the three models by various evaluation metrics can be found in the table 4.

***TableIV*.Results for evaluation of the models**

Metrics	MSE	RMSE	MAE	1-R^2
LR(One-Hot)	0.049	0.223	0.181	0.811
LR(Label)	0.058	0.241	0.203	0.947
LR(Target)	0.049	0.222	0.181	0.809
LR(WoE)	0.049	0.222	0.181	0.809
RF(One-Hot)	0.032	0.178	0.133	0.521
RF(Label)	0.033	0.181	0.135	0.535
RF(Target)	0.031	0.176	0.131	0.509
RF(WoE)	0.031	0.177	0.132	0.511
XGB(One-Hot)	0.033	0.182	0.137	0.542
XGB(Label)	0.034	0.183	0.138	0.549
XGB(Target)	0.032	0.179	0.135	0.527
XGB(WoE)	0.032	0.179	0.135	0.527

The result for the performance of 3 models by using various evaluation metrics and encoding methods can be found in Figure 1, Figure 2, Figure 3 and Figure 4.

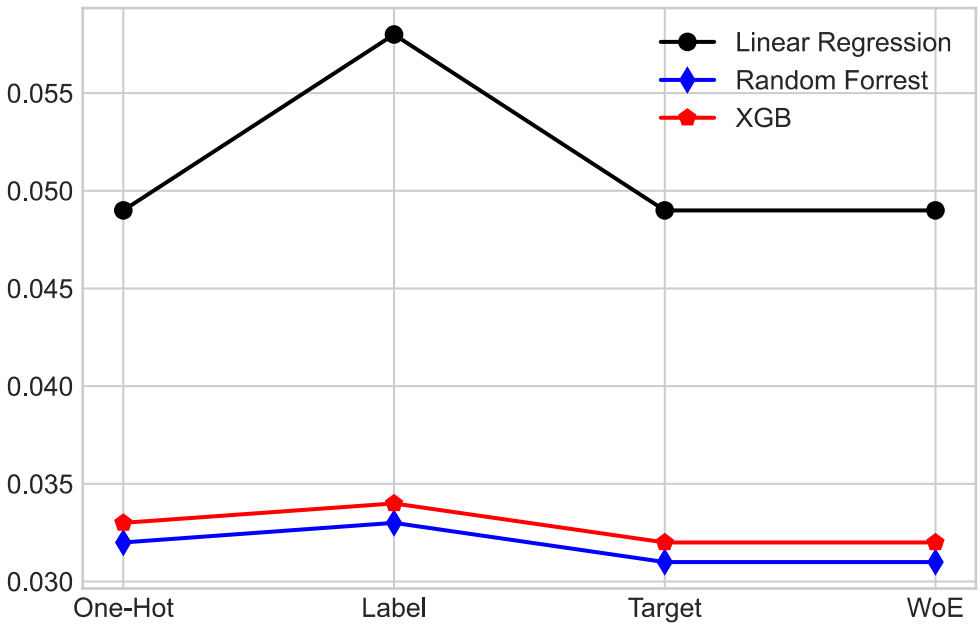

Figure 2:The MSE for 3 models by using One-Hot encoding

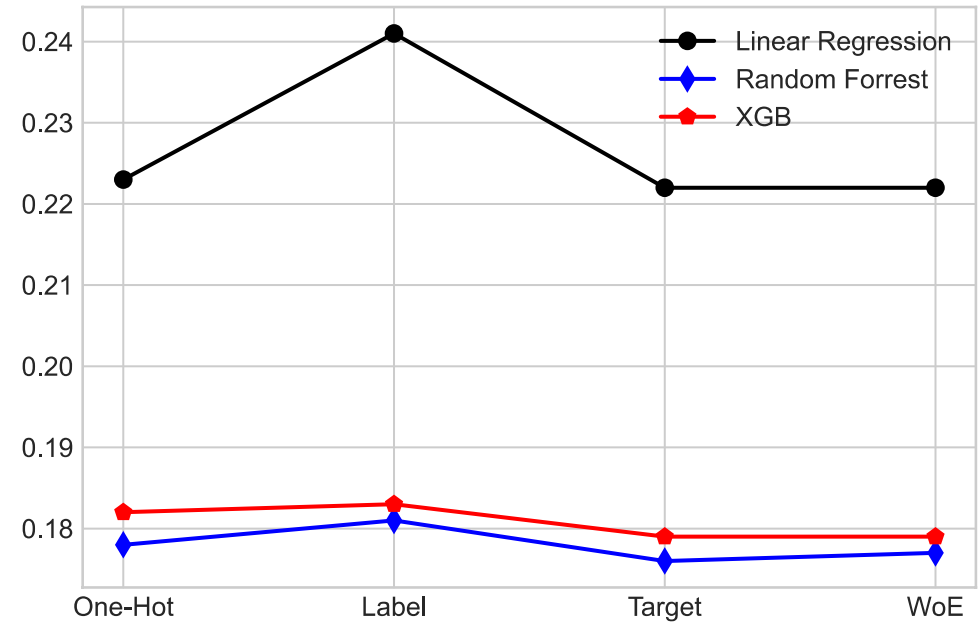

Figure 3: The RMSE for 3 models by using Label encoding

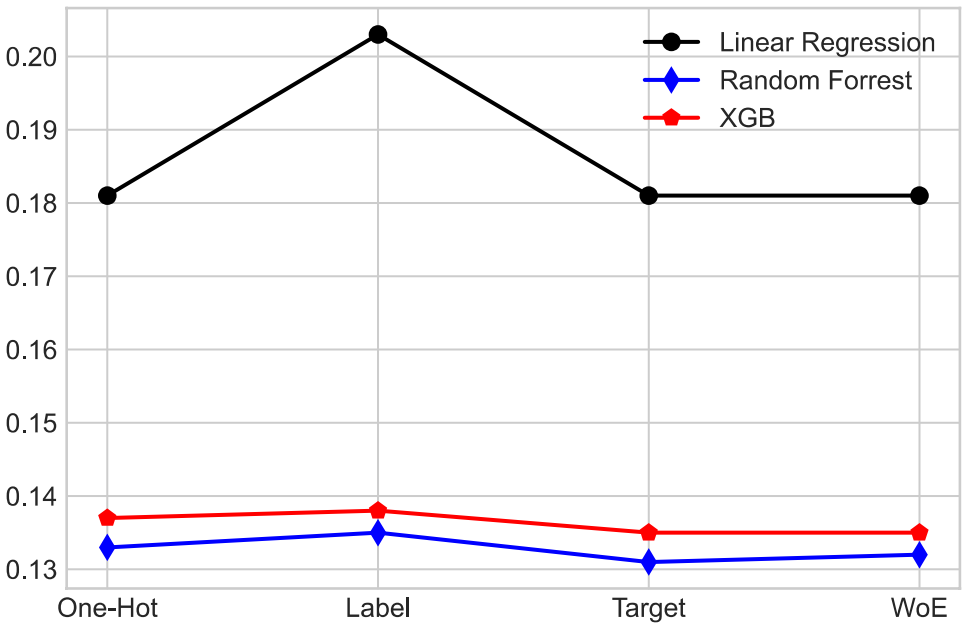

Figure 4: The MAE for 3 models by using Target encoding

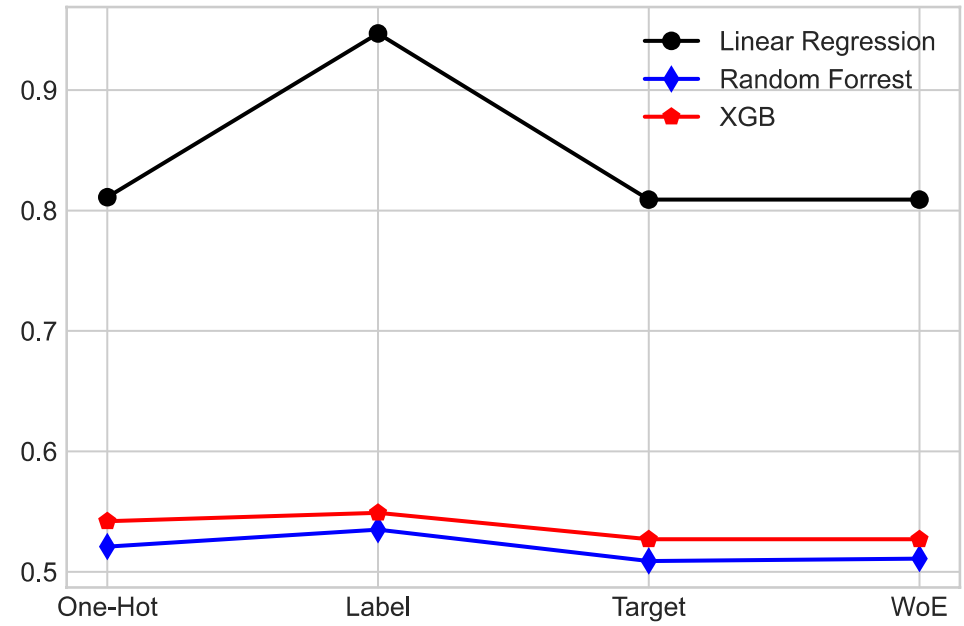

Figure 5: The 1-R^2 for 3 models by using WoE encoding

The line segments that connects the error for the linear regression indicates that it is more susceptible to the different encoding methods among which the label encoding produces the worst result. It can be also observed from the line segments that connects the error for the random forest and XGB algorithm that the random forest algorithm performs slightly better than the XGB algorithm in every encoding perspective and these two prediction models are robust to the different encoding methods.

V. CONCLUSIONS AND RECOMMENDATIONS

In this paper we have considered several models to predict the job satisfaction using the dataset retrieved from online. We conclude that the random forest model performs the best among them all. The model allows us to quantitatively predict the job satisfaction rate from the given features that are collected from the human resource department point of view. This conclusion can be useful for the study of the employee satisfaction in the management area.

REFERENCES

[1]. Yigit,I.O. and Shourabizadeh, H. (2017) An approach for predicting employee churn by using data mining. 2017 International Artificial Intelligence and Data Processing Symposium (IDAP), pp. 1-4. IEEE.
[2]. DeSousaSabbagha,M.,Ledimo,O.,andMartins,N.(2018)Predictingstaff retention from employee motivation and job satisfaction. Journal of Psychology in Africa, 28, 136-140.
[3]. Thatcher, J. B., Stepina, L. P., and Boyle, R. J. (2002) Turnover of information technology workers: Examining empirically the influence of attitudes, job characteristics, and external markets. Journal of Management Information Systems, 19, 231-261.
[4]. Eskildsen, J. K. and Dahlgaard, J. J. (2000) A causal model for employee satisfaction. Total quality management, 11, 1081-1094.
[5]. Sung, T. K., Chang, N., and Lee, G. (1999) Dynamics of modeling in data mining: interpretive approach to bankruptcy prediction. Journal of management information systems, 16, 63-85.
[6]. Ramachandran, V. and Gopal, A. (2010) Managers' judgments of performance in it services outsourcing. Journal of Management Information Systems, 26, 181-218.
[7]. Ajit, P. (2016) Prediction of employee turnover in organizations using machine learning algorithms. algorithms, 4, C5.
[8]. Zhao, Y., Hryniewicki, M. K., Cheng, F., Fu, B., and Zhu, X. (2018) Employee turnover prediction with machine learning: A reliable approach. Proceedings of SAI intelligent systems conference, pp. 737{758. Springer.
[9]. Kim,K.andLee,J.(2012)Sequentialmanifoldlearning forefficient churn prediction. Expert systems with applications, 39, 13328-13337.
[10]. Montgomery, D. C., Peck, E. A., and Vining, G. G. (2012) Introduction to linear regression analysis. John Wiley & Sons.
[11]. Liaw,A.,Wiener,M.,etal.(2002)Classification and regression by randomforest. R news, 2, 18-22.
[12]. Chen, T. and Guestrin, C. (2015) Xgboost: Reliable large-scale tree boosting system. Proceedings of the 22nd SIGKDD Conference on Know ledge Discovery and Data Mining, San Francisco, CA, USA, pp. 13-17.
[13]. https://www.kesci.com/.
[14]. https://docs.h2o.ai/h2o/latest- stable/h2o- docs/ data- munging/target- encoding.html.
[15]. https://contrib.scikit- learn.org/category_encoders/woe.html.

Haoyue Gao, et. al. "A comparison of machine learning approaches for Employee Satisfaction Prediction." *International Journal of Modern Engineering Research (IJMER)*, vol. 10(09), 2020, pp 46-53.