INTERNATIONAL JOURNAL OF

INNOVATIONS IN APPLIED SCIENCES

AND ENGINEERING

e-ISSN: 2454-9258; p-ISSN: 2454-809X

Employability of Machine Learning in the Efficacious Model Performance of Human Resource Prediction Algorithm

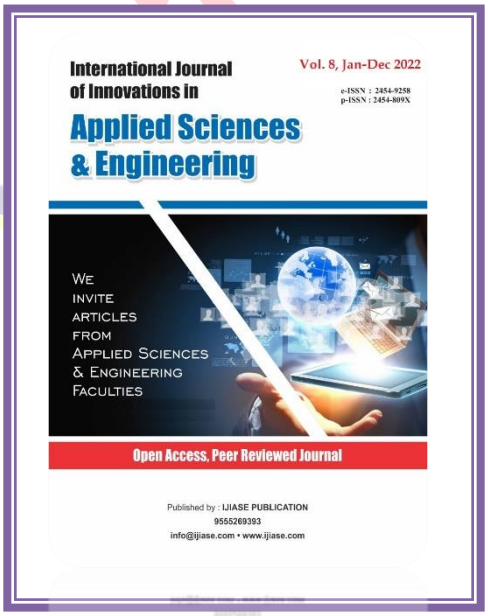
Savar Sharma
PDM University, Bahadurgarh, Haryana

Paper Received: 01st July, 2022; **Paper Accepted:** 27th August, 2022;

Paper Published: 19th September, 2022

How to cite the article:

Savar Sharma, Employability of Machine Learning in the Efficacious Model Performance of Human Resource Prediction Algorithm, IJIASE, January-December 2022, Vol 8; 76-87



ABSTRACT

A good ecological environment is crucial to attracting talents, cultivating talents, retaining talents and making talents fully effective. This study provides a solution to the current mainstream problem of how to deal with excellent employee turnover in advance, to promote the sustainable and harmonious human resources ecological environment of enterprises with a shortage of talent. This study obtains open data sets and conducts data pre-processing, model construction model optimisation, and describes a set of enterprise employee turnover prediction models based on RapidMiner workflow. The data pre-processing is completed with the help of the data statistical analysis software IBM SPSS Statistic and RapidMiner. Statistical charts, scatter plots and boxplots for analysis are generated to realise data visualisation analysis. Machine learning, model application, performance vector, and cross-validation through RapidMiner's multiple operators and workflows. Model design algorithms include support vector machines, naive Bayes, decision trees, and neural networks. Comparing the performance parameters of the algorithm model from the four aspects of accuracy, precision, recall and F1-score. It is concluded that the performance of the decision tree algorithm model is the highest. The performance evaluation results confirm the effectiveness of this model in sustainable exploring enterprise employee turnover prediction in human resource management.

INTRODUCTION

The development of technology and the increasing reliance on organizations in the Internet world have led to the growth and diversity of data [1]. Employee turnover is a focal issue in the field of organizational and human resource management research [2],[3]. But how to deal with processing using computer technology? The key factor that affects employee turnover is the importance of analysing employee turnover. It is necessary to start from the data at hand and start with various indicators of the relevant employees, such as the company's satisfaction score, working years, salary level, average monthly working hours and other indicators, to analyse and mine potential key factors, and at the same time build a predictive model for corporate employee turnover. The establishment of the model is helpful for enterprises to focus on the indicators that affect the employee turnover rate, extract the key factors of employee turnover, and explore which factors are the main factors affecting employee turnover. This is convenient for enterprises to adjust the influencing factors in a planned way, so as to manage more pertinently in their daily operations. It also analyses and guides employees who have predicted turnover tendency to enhance talent management. Data mining for business

application problems, often requires numerical prediction, precise definition of target variables and specific quantities [4]. This research dataset selects the Alibaba Tianchi public dataset[5], and uses RapidMiner software to build an employee turnover prediction model. It analyses important factors that affect employee turnover, such as evaluation, salary, overtime, etc., and it predicts whether employees have turnover intentions. The file format of this data set is CSV, and the extracted data set has a total of 5000 observation cases and 10 variable attributes.

This experiment is carried out based on determining the experimental purpose and data set. First, the data pre-processing steps and the selection of the algorithm model are carried out. At the same time, the dataset is divided into a training dataset and a test dataset. Perform parameter tuning settings for the algorithm model. After machine training, the optimal algorithm model is determined. Finally, the performance evaluation of the model is carried out. The specific experimental process is shown in Figure 1.

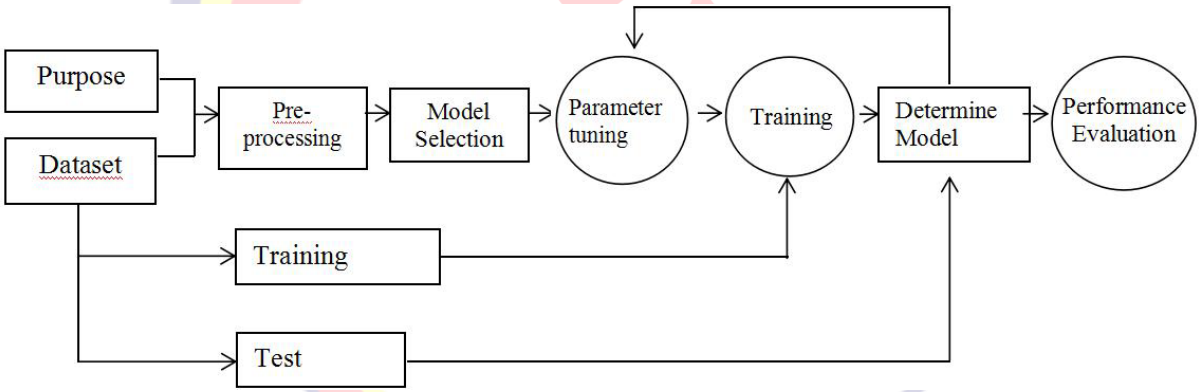


Figure 1. Detailed experimental process description.

DATA PREPROCESSING

Clear and neat data is the basis for subsequent research and analysis. This paper uses IBM SPSS Statistic and RapidMiner to complete data preprocessing, including case sampling, variable attribute transformation, correlation and rule analysis, data attribute reduction, missing value processing, and data normalization.

A. Random Sampling

Random sampling is a very important statistical analysis method, which excludes the influence and interference of human factors so that the research subjects have the

same chance to be divided into a certain treatment group. The representativeness of the sample and the balance between the sample groups are the footholds of random sampling. The application of random sampling directly affects the reliability of research results. The data is representative and the randomisation effect is good, the survey results will be more reliable, and the sampling results can be inferred to the population [6].

B. Variable

The value of the attribute variable is usually a character type or a numeric code, which cannot be applied to multi-algorithm recognition in the RapidMiner model construction below, otherwise, it will be incorrectly affected by the model construction and cause an error to be reported. Therefore, it is necessary to quantify the properties that do not meet the specification here. The non-numeric attributes are converted here by IBM SPSS Statistics. Convert the department and salary text type settings to numeric types, and re-encode the corresponding data information into different variables. When the Correlation Matrix data analysis diagram was analysed after customs clearance, the Correlation value between the independent variable last evaluation and the dependent variable left was the lowest, which was -0.017, followed by the department. Satisfaction level has the highest relationship with the dependent variable left (0.372), followed by salary, and then time_spend_company(in figure 2).

Table I. Details of Data Set Variables Attribute Conversion

Variable	Value Range	Remarks
left	0 (no),1 (yes)	dependent variable
satisfaction_level	0~1	independent variable
last_evaluation	0~1	independent variable
number_project	2~7	independent variable
average_montly_hours	96~310	independent variable
time_spend_company	2~10	independent variable
Work_accident	0 (none),1 (yes)	independent variable
promotion_last_5years	0 (no),1 (yes)	independent variable
department	Sales and so on, a total of 10	independent variable
salary	low,medium,high	independent variable

C. Association Relationship and Rule Analysis

Correlation is used to measure the direct relationship strength of each attribute variable in the existing data set, and it is a research statistical indicator. Correlation Matrix data analysis diagram can be directly obtained through Correlation Matrix Operators. Generally speaking, if the value of the relationship between two attributes is closer to 0, then the relationship between the two attributes is weaker, and vice versa, the relationship is stronger. To further confirm the relationship between attributes, it is verified by creating an association matrix, which belongs to the analysis process before attribute reduction below. The association rule operation model is established in the process of RapidMiner. Use operators such as Numerical to Binominal, FP-Growth, and Create Association to establish association rules. It is further concluded that the support degree of the last evaluation for the left is 0.239. The relationship between the last evaluation and other attributes is also low. So delete the last evaluation and do not let it participate in data model analysis.

Attributes	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	left	promotion_last_5years	department	salary
satisfaction_level	1	0.099	-0.139	-0.018	-0.087	0.068	0.372	0.006	0.014	0.054
last_evaluation	0.099	1	0.338	0.348	0.137	-0.009	-0.017	-0.013	0.004	-0.014
number_project	-0.139	0.338	1	0.412	0.195	-0.016	-0.031	0.005	-0.008	-0.011
average_montly_hours	-0.018	0.348	0.412	1	0.128	0.005	-0.082	-0.006	-0.002	-0.016
time_spend_company	-0.087	0.137	0.195	0.128	1	-0.012	-0.141	0.093	-0.055	0.055
Work_accident	0.068	-0.009	-0.016	0.005	-0.012	1	0.153	0.031	-0.016	0.031
left	0.372	-0.017	-0.031	-0.082	-0.141	0.153	1	0.058	-0.023	0.154
promotion_last_5years	0.006	-0.013	0.005	-0.006	0.093	0.031	0.058	1	-0.093	0.105
department	0.014	0.004	-0.008	-0.002	-0.055	-0.016	-0.023	-0.093	1	-0.109
salary	0.054	-0.014	-0.011	-0.016	0.055	0.031	0.154	0.105	-0.109	1

Figure 2. Correlation Matrix data analysis diagram.

D. Data Normalization

With distance-based algorithms, Normalization takes place first, scaling all attributes to the same range. The Normalization Operator performs some changes to the metabolic data set to move the employee turnover data into a specific range to make it more statistically significant. The data described below are normalized data.

MODEL CONSTRUCTION

A. Design model and Machine Learning

The prediction model belongs to the machine learning technology. It mines and processes big data, establishes an analysis model, and applies the model to make later predictions. Operations such as Filter examples, SelectAttributes, and Normalize described above will not be explained. Figure 3 and Figure 4 show the model design process of the employee turnover data mining algorithm. Figure 4 is the model nesting sub-process of Cross-Validation in Figure 3. Cross-validation will be described below.

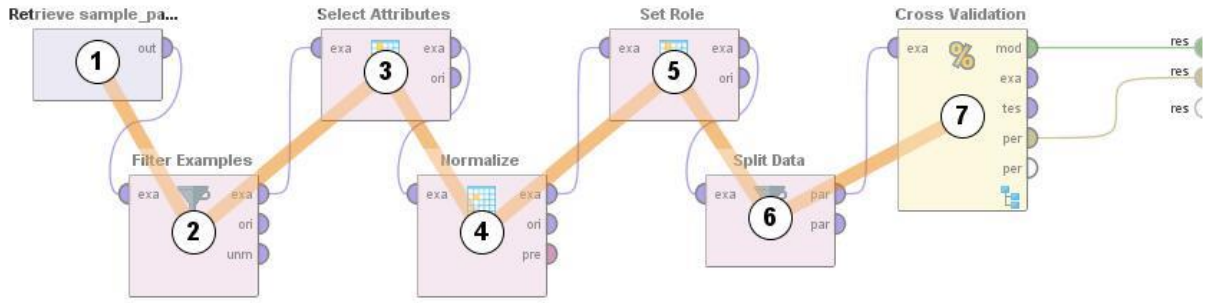


Figure 3. Design of employee turnover data mining algorithm model.

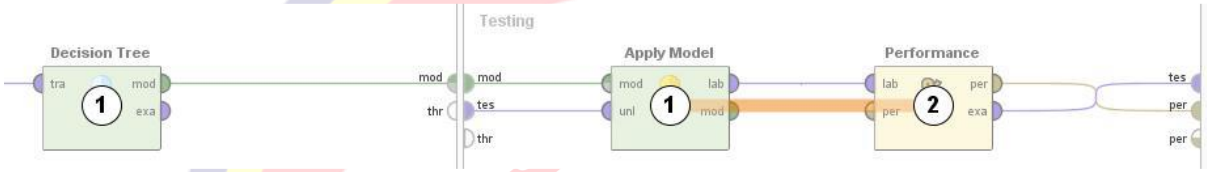


Figure 4. Design flow of cross-validation model.

1) Set predicted role

To carry out predicting, the setting of predicted roles must be completed. Set the predicted property left the role to the “label" value. In the parameters panel of "Set Role", set the attribute name to "left" and the target role to "left". In this way, the predicted attribute left has a unique predicted value, which is different from other variables. In the data statistics, the left ranks at the forefront in a significant way, which is convenient for mining and analysis.

2) Divide data

Through split data in the operator, the data set is divided into two parts: training set and test set. The test set does not participate in the training of the model. The training set accounts for 70% and the test set accounts for 30%. After dividing the data, there are 3499 employee turnover training data.

3) Machine learning and model performance

Dividing the dataset only once and then measuring the performance of the model degrades the scientific of the model due to the appearance of anomalies. At this time, cross-validation provides a method to evaluate the accuracy of the model, and cross-validation technology can cope with the processing [7]. Cross-validation repeatedly validates the measurement results based on the data set and

avoids overfitting. The cross-validation process includes two subprocesses of training the model and testing the model. In the performance parameter evaluation index, Accuracy, Precision, Recall and F1-score in studies need to be calculated [8].

The Divide data rate calculates the proportion of the total number of samples for which the prediction is correct.

Precision calculates the proportion of N returned positive samples that were predicted correctly.

Recall calculates the proportion of the number of correctly predicted positive samples among the N-returned positive samples to the true total number of positive samples.

F1-score is the harmonic mean of precision and recall. Pis precision; R is recall.

In the employee turnover data training window, the data mining algorithm successively adopts a support vector machine, naive Bayes, decision tree and neural net, and carries out machine learning and performance analysis one by one. The above algorithms are popular and concerned with predictive modelling supervision algorithms for data mining and machine learning [9],[10]. Support vector machine algorithm is a supervised algorithm for classification problems, which can deal with complex nonlinear problems; Naive Bayes is a Bayesian theorem based on probability theory, which is widely used; The decision tree contains the decision diagram and the process results of prediction, and it has the function of auxiliary decision-making; Neural network can use neuron layer to learn complex patterns, and neural network can learn the relationship between features that cannot be easily found by other algorithms[11],[12].

Refer to Table 2 for specific performance comparisons and summary results.

Table II. Comparison of Data Mining Algorithm Models

Algorithm model	Accuracy	Precision	Recall	F1-score
Support vector machine	78.11%	79.75%	95.61%	86.92%
Naive Bayes	78.94%	90.43%	80.97%	85.44%
Decision tree	97.03%	97.54%	98.61%	98.07%
neural network	93.88%	95.58%	96.47%	96.02%

4) Model performance verification and evaluation

Intuitive performance verification analysis was performed by Compare ROCs Operators in RapidMiner [13]. The X-axis is 1-specificity. The closer the X-axis is to zero, the higher the accuracy is. The Y-axis is sensitivity. The higher the Y-axis is, the higher the sensitivity is. It is obvious that the curve of the decision tree shows good performance regardless of the X-axis or Y-axis, and the ROC curve is closest to the upper left corner. In contrast, SVM ranked last. Finally, it is concluded that the performance of the employee turnover data mining algorithm model is in the order of decision tree, neural network, naïve Bayes and support vector machine.

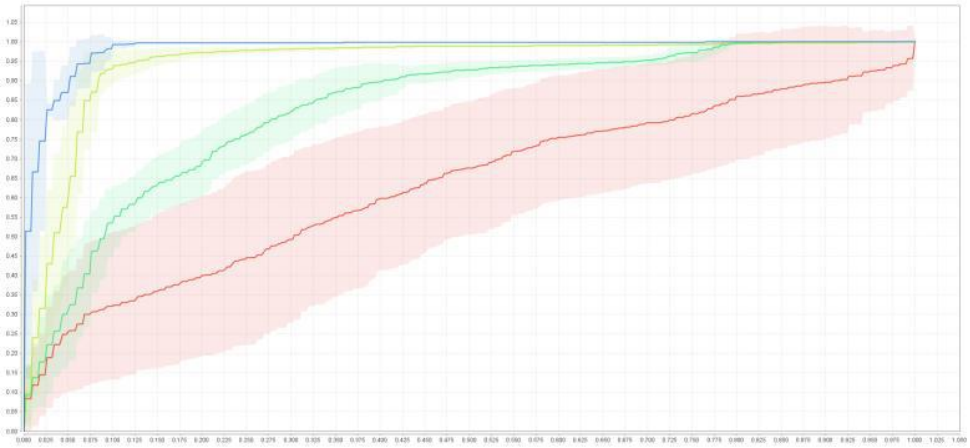


Figure 5. Comparison of ROC curve verification.

The decision tree of employee turnover data mining is shown in Figure 6. In the decision tree prediction model of employee turnover, each node in the tree corresponds to the employee attribute object, while each non-leaf node and branch corresponds to the decision-making process of the model, and the left and right branches represent the answers given by the model. Each leaf node corresponds to a predicted value represented by the path from the root to the leaf node, and the result is either 0 or 1. A decision tree is characterised by top-down determination. The higher the node, the more important its attributes are. In this model, the Satisfaction level is ranked as the root node, indicating that this attribute has the greatest impact on employee turnover in the data set. The enterprise needs to focus on how to improve employee satisfaction levels in the later stage. At the same time, the more projects, the higher the employee turnover rate. The longer the average employee works per month, the greater the turnover.

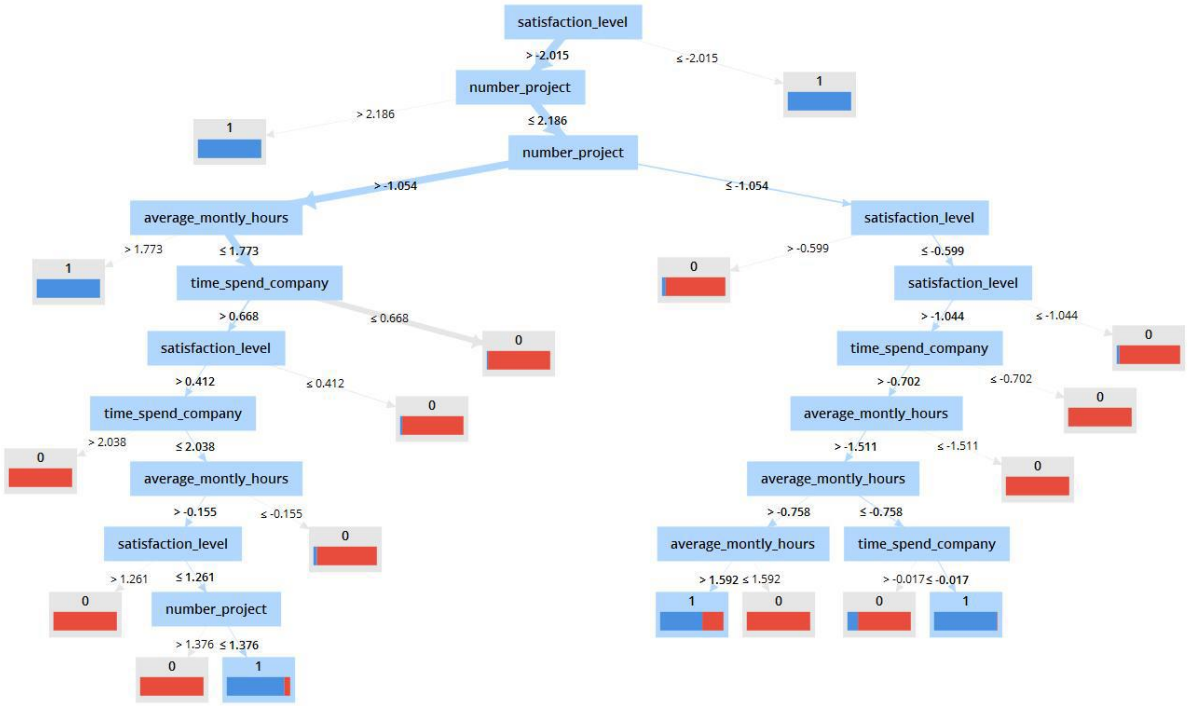


Figure 6. Decision tree of employee turnover.

CONCLUSIONS

This research strives to build a fast, convenient and rigorous scientific workflow of employee turnover prediction model, and it has been proven to be effective. The selection of data sets and model output

depends on data, and it is necessary to take into account the industry background and the influence of the times in order to get rid of the limitations of research.

REFERENCES

[1] Catia oliveira, Tiago Guimares, Filipe Portela, Manuel Santos, "Benchmarking Business Analytics Techniques in Big Data," Procedia Computer Science, vol.160, 2019,pp.690-695.

[2] Hom P W, Lee T W, Shaw J D, Hausknecht J P, "One Hundred Years of Employee Turnover Theory and Research," Journal of Applied psychology, vol.102, 2017, pp.530－545.

[3] Wang Zhongjun, Wen Lin, Long Lirong, "Boundaryless Career: Review and Prospect of Two Decades of Research," Journal of Psychological Science, vol.1, 2015, pp.243-248.

[4] Provost F., Fawcett T, "Business Problems and Data Science Solutions In Data Science for Business; Mike Loukides and Meghan Blanchette," O'Reilly Media: Sebastopol, California, USA, 2013, pp.35–42.

[5] Ali Cloud Tianchi data set. HR Analytics, 5 March 2021. [Online]. Available: https://tianchi.aliyun.com/dataset/dataDetail?dataId=92838.

[6] Li Jinchang, "Applied Sampling Technology. Science Press: Beijing, China, 2007, pp.37-38.

[7] Dossin E, Martin E, Diana P, et al.," Prediction Models of Retention Indices for Increased Confidence in Structural Elucidation during complex Matrix Analysis: Application to Gas Chromatography Coupled with High-Resolution Mass Spectrometry," Analytical Chemistry, vol. 88, 2016, pp.7539-7547.

[8] Hammad, M. A., et al., "Myocardial Infarction Detection Based on Deep Neural Network on Imbalanced Data," Multimedia Systems, vol.20,2021,p.728,.

[9] Zou Zhiwen, Zhu Jinwei, "Research and Review of Data Mining Algorithms," Computer Engineering and Design, vol.26,2005, pp. 2304-2307,.

[10] Wang Huizhong, Peng Anqun, "Existing Situation of Data Mining Research and Its Development Tendency," Industrial and Mining Automation, vol.2,2011, pp.29-32.

[11] Watson D S, Wright M N, "Testing conditional independence unsupervised learning algorithms," Machine Learning, vol.110,2021, pp.2107-2129.

[12] Prahartiwi L I, Dari W., "Komparasi Algoritma Naive Bayes, Decision Tree dan Support Vector Machine untuk Prediksi Penyakit Kanker Payudara," Journal Teknik Komputer,vol.7,2021,pp.51-54.

[13] RapidMiner Documentation, "compare_rocs. 2021.[Online]. Available:https://docs.rapidminer.com/latest/studio/operators/validation/visual/compare_rocs.html.

[14] Zhu Fei, Yue Meiqi, Zhang Jiexuan, "The mediating Role of career satisfaction and the Moderating Role of employer brand," Journal of Central University of Finance and Economics, vol.12, 2021, pp.105-118.

[15] Valentine S, Hollingworth D., "Communication of Organizational Strategy and Coordinated Decision Making as Catalysts for Enhanced Perceptions of Corporate Ethical Values in a Financial Services Company," Employee Responsibilities and Rights Journal, vol.27,2015, pp.213-229.

[16] Hina Ghous, László Kovács., "Efficiency comparison of Python and RapidMiner," Multidiszciplináris Tudományok, vol.10, 2020, pp.212-220.

[17] Moloud Abdar, "A Survey and Compare the Performance of IBMSPSS Modeler and Rapid Miner Software for Predicting Disease by Using Various Data Mining Algorithms," Cumhuriyet Science Journal, vol.36, 2015, pp.3230-3241.

[18] IBM SPSS Modeler and RapidMiner Studio. Trust Radius, December 23, 2021.[Online]. Available:https://www.trustradius.com/compareproducts/ibm-spss-modeler-vs-rapidminer-studio.