Uluslararası İleri Doğa Bilimleri ve Mühendislik Araştırmaları Dergisi
Sayı 7, S. 12-16, 9, 2023
© Telif hakkı IJANSER'e aittir
Araştırma Makalesi

IJANSER
https://as-proceeding.com/index.php/ijanser
ISSN: 2980-0811

International Journal of Advanced Natural Sciences and Engineering Researches
Volume 7, pp. 12-16, 9, 2023
Copyright © 2023 IJANSER
Research Article

Prediction of Employee Turnover with Imbalance Dataset Using Machine Learning Methods

Çetin KAYA*, Murat ŞİMŞEK [2]

[1]*Ostim Technical, University, Ankara*
[2]*Ostim Technical, University, University, Ankara*

(murat.simsek@ostimteknik.edu.tr) Email of the corresponding author

(Received: 13 September 2023, Accepted: 05 October 2023)

(3rd International Conference on Innovative Academic Studies ICIAS 2023, September 26-28, 2023)

ATIF/REFERENCE: Kaya, Ç. & Şimşek, M. (2023). Prediction of Employee Turnover with Imbalance Dataset Using Machine Learning Methods. *International Journal of Advanced Natural Sciences and Engineering Researches*, 7(9), 12-16.

Abstract **–** Employee turnover can have a significant impact on an organisation's productivity, culture and profitability. Accurately predicting employee turnover can help organisations proactively identify and address issues before they become major problems.

In this paper, various analyses were performed with the help of traditional machine learning methods using employee turnover and attrition dataset. As a result of the analyses, unbalanced data distribution was detected in the dataset. In order to solve this problem, methods for balancing up and down data sets were used. After data balancing, the k-fold method, one of the cross-validation methods, was applied to avoid overlearning. The Random Forest Classification method was selected and used together with the ROS method, which shows higher performance. GridSearchCV, a hyper-parameterisation technique, was applied to the selected model to select the best parameters. At the same time, both data pre-processing and post-processing activities were performed. As a result of the experiments conducted in the study, it was found that the data set balanced using the proposed method increased the performance values in the classification result and improved the classification performance compared to the raw data set and other sampling methods.

Keywords – Employee Turnover; Machine Learning; Imbalanced Data; Cross Validation; Classification Algorithm

I. Introduction

Employee turnover is one of the most significant issues that an organisation can face throughout its life cycle because it is difficult to predict and often creates noticeable gaps in an organisation's skilled workforce [8]. Service firms recognise that the on-time delivery of their services can be jeopardised, overall firm productivity can drop significantly and, consequently, customer loyalty can fall when employees leave unexpectedly [9]. In general, employee turnover can be divided into involuntary turnover and voluntary turnover. Involuntary turnover is usually defined as movements beyond the boundary of membership in an organisation where the employee leaves only voluntarily. It exhibits mild emotions. On the other hand, voluntary turnover, movements beyond the boundary of a person's membership within an organisation, on which the employee harbours heavy emotions [2].

In real data, there are many cases where the number of instances in one class is much less than the number of instances in another class. The instances in other classes are referred to as an unbalanced dataset.[1] We have observed some interesting trends/results based on our examination of the studied artefacts and some of the key findings are summarised below. Among the Data Level methods, experimental results of related studies generally show that Random Oversampling (ROS) gives better classification performance than Random Under-Sampling or Synthetic Minority Oversampling Technique (SMOTE). [4] A natural approach to overcome this class imbalance problem is to rebalance the training target.[3] Many methods are used for this case.

II. Materials and Method

Describe in detail the materials and methods used when conducting the study. The citations you make from different sources must be given and referenced in references.

A. *Introduction of the Data Set*

In this case study, an HR dataset is taken from IBM HR Analytics Employee Attrition and Performance which contains employee data of 1,470 employees along with various information about the employees. I will use this dataset to predict when employees will quit by understanding the main causes of employee attrition. As stated on the IBM website "This is a fictional dataset created by IBM data scientists". Its main purpose was to demonstrate the IBM Watson Analytics tool for employee attrition. The dataset consists of 35 variables. 26 of these variables are numeric and 9 are categorical variables.

B. *Classification Methods*

In this paper, I have applied the most widely used machine learning methods in the supervised classification category. The machine learning algorithms used are visualised in Figure 1.

In this case study, an HR dataset was taken from IBM HR Analytics Employee Attrition and Performance, which contains employee data of 1,470 employees along with various information about the employees. I will use this dataset to predict when employees will quit by understanding the main causes of employee attrition. As stated on the IBM website "This is a fictional dataset created by IBM data scientists". Its main purpose was to demonstrate the IBM Watson Analytics tool for employee attrition. The dataset consists of 35 variables. 26 of these variables are numeric and 9 are categorical variables.

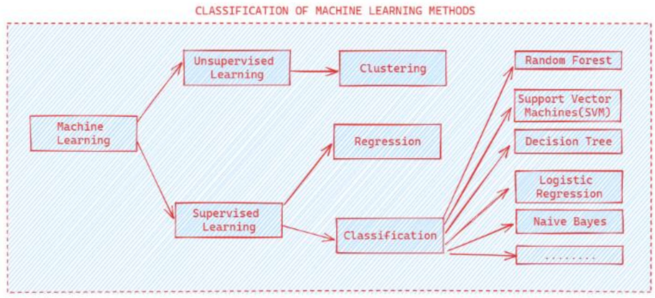

Fig 1. Machine Learning Methods Grouping Machine Learning Methods Grouping

Machine Learning is a sub-major branch of artificial intelligence that involves developing algorithms and models that have the ability to analyse collected data and gain the ability to learn from this data, measure and improve the performance of machines on a specific task or set of tasks without programming. Supervised Learning involves training models on labelled data where the correct output or outcome is already known. The most commonly used classification algorithms are described below respectively. Random Forest is a machine learning algorithm used for classification, regression and other tasks, based on the idea of building multiple decision trees and combining the outputs to make a final prediction. Support Vector Machine (SVM) is a popular machine learning algorithm used for classification and regression analysis. Decision Tree is a type of machine learning algorithm used for classification and regression problems. Logistic Regression is a statistical algorithm used in binary classification problems. It is a type of regression algorithm that models the probability of a binary response variable (0,1) based on one or more predictor variables. K-nearest neighbour algorithm is a non-parametric, supervised learning classifier that uses proximity to make classifications or predictions about the grouping of an individual data point.

C. *Evaluation Metrics*

In machine learning, the Confusion (Error) Matrix in Figure 2 is used to measure and evaluate the performance of classification models by comparing the predicted values with the actual values.

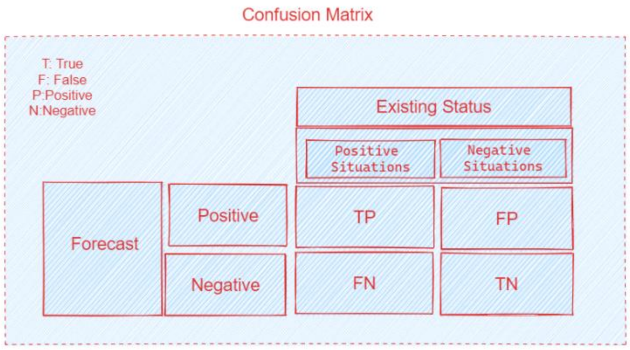

Fig 2. Trained models evaluation metrics

Accuracy: How many of all classes (positive and negative) we predicted correctly.

Precision: It can be explained by saying how many of all the classes we predicted as positive are actually positive. Equation (1) is shown below.

$$Precision = \frac{TP}{FP+TP} \tag{1}$$

Recall: It can be explained by saying how many of all positive classes we guessed correctly. Equation (2) is shown below.

$$Recall = \frac{TP}{TP+FN} \tag{2}$$

F-Score: It is difficult to compare two models with low sensitivity and high recall or vice versa. That is why we use F-Score to make them comparable. F-score helps to measure recall and precision at the same time. It uses the Harmonic Mean instead of the Arithmetic Mean, penalising outliers more. It is shown in equation (3).

$$F - measure = 2x \frac{Precision}{Recall+Precision} \tag{3}$$

D. *Oversampling Methods for Unbalanced Data Set*

Oversampling Methods for Unbalanced Data SetFor many years, there has been extensive research on the use of oversampling techniques to address class imbalance. Extended research on this technique avoids the loss of important information while preserving the original data set. This technique creates a balanced data set by generating new samples that should be added to the minority sample class. Oversampling can be done either through random oversampling. The data set is balanced either by replicating the existing minority sample class or by synthetic oversampling. The data set is balanced by creating new synthetic minority samples through linear interpolation.[2] Applying resampling strategies to achieve a more balanced data distribution is an effective solution to the imbalance problem.

The data balancing methods used in this research are Random Oversampling Samples (ROS), Synthetic Minority Oversampling Technique (SMOTE), Random Subsampling (RUS), Near Miss methods. Data were balanced by down and up sampling. ROS is the simplest and oldest method of dealing with this problem. In this method, on the one hand, the classifier is trained until it reaches the desired ratio, on the other hand, it is used to balance the distribution by randomly copying the minority class samples to bring them closer to the larger class [10]. SMOTE, this technique is one of the most frequently used methods in imbalanced data balancing. It aims to balance the class distribution by replicating the minority class samples and randomly increasing this replication activity. It can be grouped under high sampling algorithms. RUS is a quick and easy way to balance the data by randomly selecting a subset of data for the intended majority data classes. It can be grouped under the low sampling class. Near Miss is an algorithm that can help to balance an unbalanced dataset. It can be grouped under downward, i.e. undersampling algorithms and is an effective way to balance data.

III. FINDINGS AND INTERPRETATIONFINDINGS AND INTERPRETATION

In this research, we have implemented an end-to-end machine learning model from data preprocessing, to measuring the performance of the model, to subsequent cross-validation.

For feature selection, we analysed the correlation between the data using the heat map method as shown in Figure 4. With the help of this analysis, the variables with high correlation were selected as single variables.

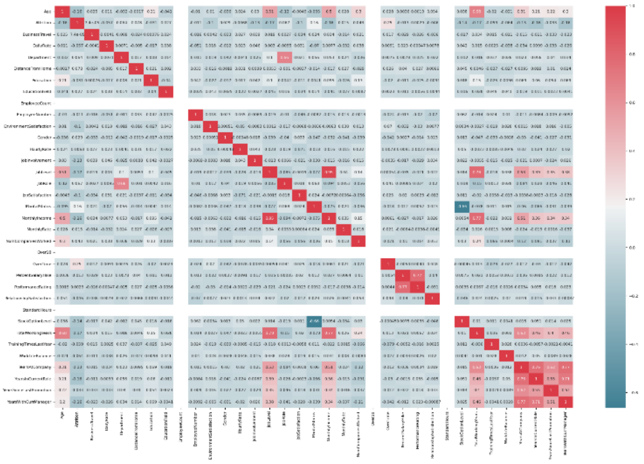

Fig. 3 Correlation between features

Firstly, with the help of the model selection method, the data were divided into two as training and test and made ready for use. Logistic Regression, Support Vector Machine (SVM), Nearest Neighbour (KNN), Decision Trees, Random Forest Classification algorithms were applied to the data set respectively. The results are shown in Table 1 below.

Table 1.Performance metrics of traditional Machine Learning algorithms

ML Algoritm	Model Accuracy	Precision 0	Precision 1	Recall 0	Recall 1	F1-Score 0	F1-Score 1
Logistic regression	0.71	0.88	0.68	0.97	0.32	0.92	0.43
SVM	0.80	0.92	1.00	0.15	0.92	0.26	0.86
KNN	0.81	0.85	0.45	0.99	0.06	0.91	0.11
Decision Tree	0.54	0.85	0.25	0.83	0.29	0.84	0.27
Random Forest	0.52	0.85	0.64	0.99	0.11	0.92	0.11
Naive Bayes	0.74	0.87	0.45	0.93	0.36	0.90	0.36

The performance results show that there is an unbalanced data distribution problem. In order to eliminate this problem, ROS, SMOTE, RUS, Near Miss methods were applied for upstream and downstream data set balancing.

The results obtained after the application of Logistic Regression and data balancing methods are shown in Table 2 below.

Table 2. Logistic regression processing methods with unbalanced data methods

ML Algoritm	Model Accuracy	Precision 0	Precision 1	Recall 0	Recall 1	F1-Score 0	F1-Score 1
ROS-Logistic regression	0.72	0.88	0.81	0.79	0.89	0.83	0.85
SMOTE-Logistic regression	0.73	0.86	0.86	0.86	0.86	0.86	0.86
RUS-Logistic regression	0.70	0.71	0.73	0.73	0.70	0.72	0.72
Near Miss-Decision Tree	0.70	0.69	0.77	0.81	0.63	0.74	0.69

The results obtained after the application of Support Vector Machine (SVM) and data balancing methods are shown in Table 3 below.

Table 3. Support Vector Machine(SVM) processing methods with unbalanced data methods

ML Algoritm	Model Accuracy	Precision 0	Precision 1	Recall 0	Recall 1	F1-Score 0	F1-Score 1
ROS-SVM	0.84	0.88	0.81	0.79	0.89	0.83	0.85
SMOTE-SVM	0.86	0.86	0.86	0.86	0.86	0.86	0.86
RUS-SVM	0.72	0.73	0.73	0.70	0.72	0.72	0.73
Near Miss-SVM	0.72	0.77	0.81	0.63	0.74	0.69	0.77

The results obtained after the application of data balancing methods with Nearest Neighbour (KNN) are shown in Table 4 below.

Table 4. Nearest Neighbour (KNN) processing methods with unbalanced data methods

ML Algoritm	Model Accuracy	Precision 0	Precision 1	Recall 0	Recall 1	F1-Score 0	F1-Score 1
ROS-KNN	0.82	0.90	0.77	0.72	0.92	0.80	0.84
SMOTE-KNN	0.79	0.99	0.71	0.59	0.99	0.74	0.83
RUS-KNN	0.65	0.63	0.69	0.74	0.57	0.69	0.62
Near Miss-KNN	0.72	0.56	0.66	0.81	0.38	0.67	0.48

The results obtained after the application of Decision Trees and data balancing methods are shown in Table 5 below.

Table 5. Decision Trees processing methods with unbalanced data methods

ML Algoritm	Model Accuracy	Precision 0	Precision 1	Recall 0	Recall 1	F1-Score 0	F1-Score 1
ROS-Decision Trees	0.74	1.00	0.65	0.47	1.00	0.64	0.79
SMOTE-Decision Trees	0.67	0.83	0.63	0.49	0.87	0.61	0.73
RUS-Decision Trees	0.50	0.52	0.54	0.70	0.35	0.59	0.43
Near Miss-Decision Trees	0.51	0.50	0.50	0.59	0.41	0.54	0.45

The results obtained after the application of Random Forest Algorithm and data balancing methods are shown in Table 6 below.

Table 6. Random Forest processing methods with unbalanced data methods

ML Algoritm	Model Accuracy	Precision 0	Precision 1	Recall 0	Recall 1	F1-Score 0	F1-Score 1
ROS-Random Forest	0.80	1.00	0.70	0.57	1.00	0.73	0.82
SMOTE-Random Forest	0.86	0.91	0.85	0.84	0.92	0.87	0.88
RUS-Random Forest	0.58	0.58	0.64	0.74	0.45	0.65	0.53
Near Miss-Random Forest	0.60	0.60	0.67	0.76	0.49	0.67	0.56

As a result of the experiments conducted in this research, ROS was chosen as the unbalanced data method and Random Forest Model as the classification model. In order to avoid overlearning, k-Fold method was also applied as a cross-validation method. The performance of the prediction classification algorithms with unbalanced employee loss data was improved by using ROS, GridSearchCV, Random Forest Model and Hiperparametre tuning methods as shown in Table 8.

Table 7. Improvement results of processing unbalanced data with random forest, ros and gridsearchCv methods

ML Algoritm	Model Accuracy	Precision 0	Precision 1	Recall 0	Recall 1	F1-Score 0	F1-Score 1
ROS-Random Forest	0.96	0.97	0.94	0.94	0.98	0.96	0.96

IV. CONCLUSION

In this paper, methods for processing the effects of imbalanced data on machine learning algorithms are investigated. [6] An HR dataset containing data of 1,470 employees is provided. This dataset was used to predict when employees will leave their jobs by understanding the main causes of employee turnover. Traditional machine learning algorithms were trained with 75% of the data and tested with 25% of the data. [6] In this unbalanced data study, turnover prediction was performed by considering yes and no rates depending on the attrition variable. Balancing methods were performed by using data downbalancing and upbalancing techniques for the insufficient no data set. Random forest classification values showing high F1-score values in the first stage are 0.92 and 0.19. After applying the data balancing methods, the highest performance values are provided by Random Forest Algorithm, ROS, GridSearchCV methods and the F1-Score values are obtained as "0.96" and "0.96".

V. REFERENCES

[1] D. Ramyachitra and P. Manikandan," IMBALANCED DATASET CLASSIFICATION AND SOLUTIONS: A REVIEW", International Journal of Computing and Business Research (IJCBR), vol. 5, Issue 4, July 2014

[2] Hong, W.-C., Pai, P.-F., Huang, Y.-Y., & Yang, S.-L. (2005). Application of Support Vector Machines in Predicting Employee Turnover Based on Job Performance. Advances in Natural Computation, 668–674. doi:10.1007/11539087_85

[3] Danquah, R. A., Handling Imbalanced Data: A Case Study For Binary Class Problems, Department of Mathematics Southern Illinois University Edwardsville, IL 62026

[4] Kim, J., Jeong, J., & Shin, J. (2020). M2m: Imbalanced Classification via Major-to-Minor Translation. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). doi:10.1109/cvpr42600.2020.01391

[5] Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A and Seliya, N., A survey on addressing high-class imbalance in big data, Leevy et al. J Big Data (2018) 5:42

[6] Şimşek, M., Daş, A. S. ,The Effect of Handling Imbalanced Datasets Methods on Prediction of Entrepreneurial Competency in University Students, 2022, www.iceans.org

[7] Zhao, Y., Hryniewicki, M. K., Cheng, F., Fu, B., & Zhu, X. (2018). Employee Turnover Prediction with Machine Learning: A Reliable Approach. Intelligent Systems and Applications, 737–758. doi:10.1007/978-3-030-01057-7_56

[8] Alao, D., Adeyemo, A.B.: Analyzing employee attrition using decision tree algorithms. Comput. Inf. Syst. Dev. Inform. Allied Res. J. 4 (2013)

[9] Sexton, R.S., McMurtrey, S., Michalopoulos, J.O., Smith, A.M.: Employee turnover: aneural network solution. Comput. Oper. Res. 32, 2635–2651 (2005)

[10] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, Learning from imbalanced data sets. Springer, 2018.