Advances and Applications in Mathematical Sciences
Volume 21, Issue 3, January 2022, Pages 1479-1489

PREDICTION OF ABSENTEEISM AT WORK WITH MULTINOMIAL LOGISTIC REGRESSION MODEL

DEMUDU NAGANAIDU, ZARINA MOHD KHALID and SURESH GOVINDAN

Center for Postgraduate Studies
Asia Metropolitan University
81750 Johor Bahru, Johor, Malaysia
E-mail: demudu@amu.edu.my

Department of Mathematical Sciences
Faculty of Science 81310
UTM Johor Bahru, Johor, Malaysia
E-mail: zarinamkhalid@utm.my

Professor of Mathematics
Sanskrithi School of Engineering
Puttaparthi, Andhra Pradesh-515134, India
E-mail: suresh.g@sseptp.org

Abstract

Employees who are absent from work and their job responsibilities cause critical problems in the employee-employer relationship. An employee's absence from work may appear inconsequential to them. However, absenteeism may impact some organizations a great deal of cost in terms of lost productivity. This study investigates an example of absenteeism at a Brazilian courier company and the underlying reasons for absence. The objective of this paper is to use the Multinomial Logistic Regression model to predict 3 classes of employee absence. True Positive, True Negative, False Positive, False Negative, Precision, Recall and F1-Score measures were used to assess model accuracy. The article reports Multinomial Logistic Regression can predict with an accuracy of 88% without balance classes but failed to predict Class 3 completely with zero F1 Score. Accuracy reduced to 78% upon balance classes introduced with Synthetic Minority Oversampling Technique (SMOTE) but improved the F1 score for predicting Class 3.

2020 Mathematics Subject Classification: Primary 05A15; Secondary 11B68, 34A05.
Keywords: Absenteeism; classification; multinomial logistic regression; machine learning; feature selection.
Received October 31, 2021; Accepted November 10, 2021

1. Introduction

Organizational objectives only are met if employees show up and do so on time. Employees who are absent from work and their job responsibilities cause critical problems in the employee-employer relationship. An employee's absence from work may appear inconsequential to them. However, absenteeism may impact some organizations a great deal of cost in terms of lost productivity.

According to Badubi [1], some of the factors contributing to absenteeism include family responsibilities, pregnancy and maternity leaves, minor illness, acute medical conditions, injuries, stress, burnout and fatigue, alcohol or drug-related conditions, bad weather and transport problem, etc.

Understanding employee absenteeism is one of the key tasks of any Human Resources department (HR) of any organization. High employee absenteeism is important indicator for low employee motivation. In a study by ten Brummelhuis et al. [2] a co-worker of an absenteeism-prone employee is more likely to call in sick.

This study uses a Multinomial Logistic Regression model (MLR) to predict absenteeism at work based on data collected by Martiniano et al. [3] for a courier service company in Brazil. The rest of the article is organized as follows: In Section 2, recent related works are presented. In Section 3, the research methodology is outlined. Section 4 explains about dataset. Section 5 is the results and conclusion in Section 6.

2. Related Works

Several classification models were applied to the dataset [3] to predict absenteeism and to understand the underlying factors. Al-Rasheed [4] used seven (7) different classification models: Naive Bayes, Logistic Regression, Multilayer Perceptron, K-Nearest Neighbour, Bagging, J48, and Random Forest. To select the most important dependant variables, three (3) features selection algorithms, Relief-based feature selection (RFS), Correlation-based feature selection (CFS) and Information-gain feature selection (IGFS) were applied. Based on performance metrics, bagging classification model was selected as the best with 92% accuracy. Ayman Al-Zibdeh et al., [5] fitted

Artificial Neural Network model with all nineteen (19) independent variables, yielding a 99% accuracy.

Skorikov et al. [6] performed prediction with four (4) classification models on the same dataset, namely Zero R., Naïve Bayes and K-Nearest Neighbour (KNN). The dependent variable converted to three (3) classes: 0 hours as Class A, 1-15 hours Class B and 16-120 hours as Class C. Feature selection, CFS algorithm utilised to select the dependent variables. Based on the CFS ranking, three (3) experiments were conducted: Experiment A, Experiment B, and Experiment C. Experiment A consist of four (4) independent variables ('Month of absence', 'Age', 'Disciplinary failure', 'Social drinker'), Experiment B with all nineteen (19) independent variables while Experiment C with single independent variable i.e. 'Disciplinary failure'. Only Experiment C is not inclusive of the two variables, 'Reason for absence' and 'Month for absence'. The dependent variable created with three (3) classes however was imbalance. To address the imbalance classes problem, Synthetic Minority Oversampling Technique (SMOTE) was applied. The author concluded model can accurately predict absenteeism with over 92% accuracy.

Ali Shah et al., [7] built a Deep Neural Network (DNN) to predict absenteeism and compared it to a Shallow Neural Network, as well as Decision Tree, Support Vector Machine (SVM), and Random Forest models. All the nineteen (19) variables were used during the modelling fitting process. In conclusion, it was found that DNN is the best model for predicting work absenteeism. Wahid et al., [8] used all twenty (20) attributes on four (4) classification models: Decision Tree, Gradient Boosted Tree, Random Forest and Tree Ensemble. Absenteeism variable categorised to four (4) classes: 0 hours-Not Absent, 1-7- Hours, 8-39-Days and above 40 hours-Weeks. Gradient Boosted Tree produced the highest accuracy of 82%.

One mistake in past research [4]-[8] was the use of a feature selection algorithm on all independent variables. Two variables, 'Reason for absence' and 'Month for absence' recorded zero (0) values for employees with no absence. Both variables already have information on absence and non-absence. Hence including both variables into the model fitting, leads to the high accuracy of prediction. Although the two variables have been excluded in experiment C by Skorikov et al., [6], the MLR model was omitted. Hence, the mistake is addressed in this study by fitting the MLR model by excluding the

two variables, Reason for absence and Month for absence, and evaluating the prediction performance.

3. Research Methodology

A classification with MLR model is proposed in this study to predict absenteeism in a courier service company. Dataset was pre-processed before further analysis and training the model. To train the model, Scikit-Learn a Python package [9] for data science, was utilized. Performance of the MLR is compared two other models namely, Decision Tree (Tree) and K-Nearest Neighbour (KNN). The research framework is outlined in Figure 1.

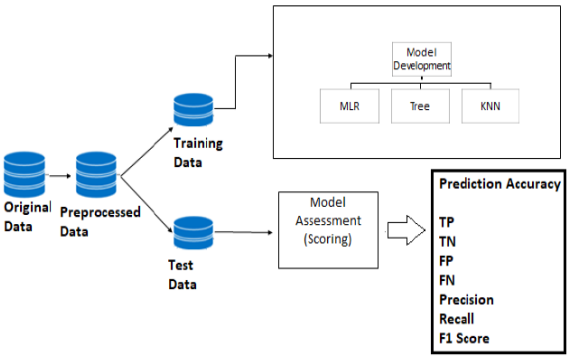

Figure 1. Research Framework.

Model accuracy assessed with the commonly used evaluation metrics [5], [6], [8], True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). For three (3) class prediction problem the confusion matrix shown in Figure 2. For Class 1, $TP = N11$, $TN = N22 + N23 + N32 + N33$, $FP = N21 + N31$ and $FN = N12 + N13$. Similarly, the same can be computed for Class 2 and Class 3. Precision, Recall and F1-Score computed from classification report as summarised in Table I. Both evaluation metrics are available from Scikit-Learn library [9].

		Predicted Class		
		1	2	3
Actual Class	1	N11	N12	N13
	2	N21	N22	N23
	3	N31	N32	N33

Figure 2. Confusion Matrix.

Table I. Classification Report.

Precision-Accuracy of positive predictions.	Precision $= TP/(TP + FP)$
Recall: Fraction of positives that were correctly identified.	Recall $= TP/(TP + FN)$
F1 score-Percentage of positive predictions were correct.	F1 Score = 2*(Recall * Precision) / (Recall + Precision)

4. Dataset

The dataset [3], which spans three years was collected between July 2007 and July 2010, is publicly available in UCI Machine Learning website. The dataset consists of 740 records with 22 attributes including 'ID' as the identification of the employee and 'Absenteeism time in hours' being the dependent variable. The rest of attributes identified as independent variables are as per Table II. International Code of Diseases (ICD) stratified into 21 categories (I to XXI) and 7 categories without ICD codes recorded to study the impact of various health-related issues to absenteeism.

The original dataset required pre-processing before models training. Variable 'ID' dropped as it is a unique number for each employee and has no use in prediction models. Figure 3 shows the distribution of dependent variable, absenteeism in hours. According to Skorikov et al. [6] the variable absenteeism implies the presence of three classes. Hence absenteeism variable categorised into three classes as per Table III. Figure 4 shows the

number of records in each class. Class 1 is 44 or 5.95% records. Class 2 is 633 or 85.54% and Class 2 is 63 or 8.51%.

Both variables 'Reason for absence' and 'Month of absence' are inclusive of information on non-absence, thus excluded in all model development. Other categorical variables in Table II with nominal scale converted into dummy variables. Numerical variables standardised or normalised due to different measurement scales. After the data pre-processing, it was split into training and testing data in the ratio of 80:20. This results in 592 records were in training data while 148 records in testing data.

Table II. Independent variables.

Categorical	Numerical
'Reason for absence'	'Transportation expense'
'Month of absence'	'Distance from Residence to Work'
'Day of the week'	'Service time'
'Seasons'	'Age'
'Disciplinary failure'	'Workload Average/day'
'Education'	'Hit target'
'Social drinker'	'Son'
'Social smoker'	'Pet'
	'Weight'
	'Height'
	'Body mass index'

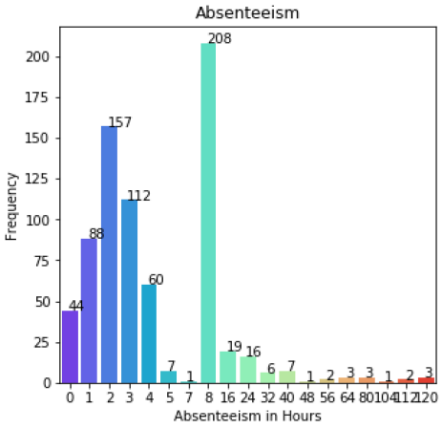

Figure 3. Distribution of Absenteeism.

Table III. Absenteeism Category Classes.

Absenteeism in hours (time)	Class
0	1
1-15	2
16-120	3

5. Results and Discussions

The training data fitted to the MLR, Tree, and KNN models. The result summarized in Table IV. To deal with the imbalance dataset SMOTE [10] technique applied to produced balance dataset. The balance dataset refitted to all three (3) models. The results in Table V. Average prediction accuracy (Accuracy) for each model was achieved by repeated 5-fold cross-validation.

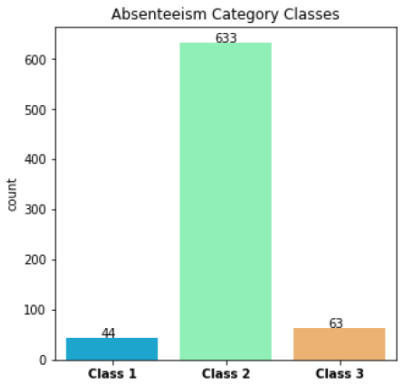

Figure 4. Absenteeism Class Distribution.

From Table IV it can been seen that MLR predict with highest accuracy of 88% compared to other models but failed to detect Class 3 completely with zero F1 score. KNN model recorded 85% accuracy but zero F1 score for Class 1 and 3. Tree model prediction accuracy is 74% and F1 score of 10% for Class 3. Upon introducing balanced classes dataset with SMOTE, the F1 Score for Class 3 for MLR and KNN improved to 20% and 23% respectively. Tree model accuracy rate improved to 86%, however recorded lower F1 score for Class 3, i.e. from 10% reduced to 9%. Overall KNN recorded lower F1 score for Class 1 and Class 2 compared to MLR and Tree.

Although the Tree model has higher accuracy than the MLR model, it involves nodes, and it can take a lot of mental effort to understand all the splits that lead up to a given prediction. Decision tree based on a balance dataset is shown Figure 7. An MLR model, on the other hand, is only a list of coefficients. In the interest of space, only three variables, namely 'Social drinker', 'Social smoker' and 'Disciplinary failure' coefficients, are shown in Table VI. Among the three variables, the 'Disciplinary failure' weight is much higher than the other two variables. Thus, this variable is vital in predicting absenteeism.

Table IV. Results Class.

		Class		
		1	2	3
	Precision	1	0.89	0

	Recall	0.91	1	0
	F1-Score	0.95	0.94	0
	Accuracy	0.88		
Tree	Precision	1	0.89	0.14
	Recall	0.91	0.95	0.07
	F1-Score	0.95	0.92	0.10
	Accuracy	0.74		
KNN	Precision	0	0.83	0
	Recall	0	1.00	0
	F1-Score	0	0.91	0
	Accuracy	0.85		

Based on full analysis, variables 'Disciplinary failure', 'Weight' and 'Body mass index' are key attributes in determining the absenteeism of an employee for three classes of absenteeism.

Table V. Results with Balance Classes.

		Class		
		1	2	3
MLR	Precision	0.91	0.73	0.14
	Recall	0.91	0.81	0.36
	F1-Score	0.91	0.90	0.20
	Accuracy	0.78		
Tree	Precision	1	0.89	0.12
	Recall	0.91	0.94	0.07
	F1-Score	0.95	0.92	0.09
	Accuracy	0.87		

KNN	Precision	0.44	0.89	0.16
	Recall	0.36	0.73	0.43
	F1-Score	0.40	0.80	0.23
	Accuracy	0.84		

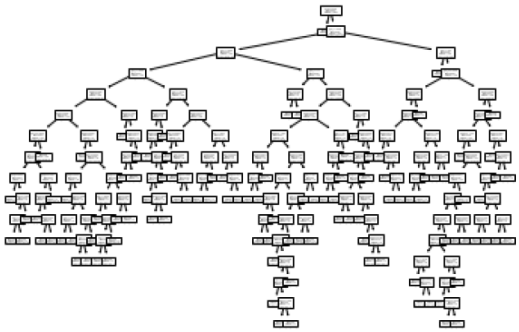

Figure 7. Decision Tree.

Table VI. Coefficients Values of Selected Variables.

	Class		
Variables	1	2	3
Social drinker	0.2087	-0.7321	0.5234
Social smoker	0.2844	0.3855	-0.6699
Disciplinary failure	6.7237	-3.4327	-3.2910

6. Conclusion

In this study absenteeism dataset [3] analysed. Three (3) models developed for prediction of three (3) types of absenteeism classes. It has been found that MLR model accuracy was lower compared to Tree model, but easier to comprehend. For HR managers who are interested to address the root cause of the absenteeism besides predicting the absenteeism among the employees, MLR models can be utilised to identify the factors contributing high absenteeism and appropriate action can be taken to reduce the absenteeism.

References

[1] R. M. Badubi, A critical risk analysis of absenteeism in the work place, J. Int. Bus. Res. Mark. 2(6) (2017), 32-36, doi: 10.18775/jibrm.1849-8558.2015.26.3004.

[2] L. L. ten Brummelhuis, G. Johns, B. J. Lyons and C. L. Ter Hoeven, Why and when do employees imitate the absenteeism of co-workers?, Organ. Behav. Hum. Decis. Process. 134 (2016), 16-30, doi: 10.1016/j.obhdp.2016.04.001.

[3] A. Martiniano, R. P. Ferreira, R. J. Sassi and C. Affonso, Application of a neuro fuzzy network in prediction of absenteeism at work, Iber. Conf. Inf. Syst. Technol. Cist. (2012), 1-4.

[4] A. Al-Rasheed, Identification of important features and data mining classification techniques in predicting employee absenteeism at work, Int. J. Electr. Comput. Eng. 11(5) (2021), 4587-4596, doi: 10.11591/ijece.v11i5

[5] A. Ayman Al-Zibdeh, R. Adnan Abu Hassanein, S. Ahmed Al-Qassas and F. Naji Abu Tir, Workplace Absenteeism Prediction using ANN, Int. J. Acad. Inf. Syst. Res. 5(1) (2021), 41-47, [Online]. Available: https://archive.ics.uci.edu/ml/datasets.html.

[6] M. Skorikov et al., Prediction of Absenteeism at Work using Data Mining Techniques, Proc. ICITR - 5th Int. Conf. Inf. Technol. Res. Towar. New Digit. Enlight., 2020, doi: 10.1109/ICITR51448.2020.9310913.

[7] S. A. Ali Shah, I. Uddin, F. Aziz, S. Ahmad, M. A. Al-Khasawneh and M. Sharaf, An Enhanced Deep Neural Network for Predicting Workplace Absenteeism, Complexity 2020, doi: 10.1155/2020/5843932.

[8] Z. Wahid, A. K. M. Z. Satter, A. Al Imran and T. Bhuiyan, Predicting absenteeism at work using tree-based learners, Pervasive Health Pervasive Comput. Technol. Healthc. (2019), 7-11, doi: 10.1145/3310986.3310994.

[9] P. Fabian, V. Gael, G. Alexandre, M. Vincent, T. Bertrand and G. Olivier, Scikit-learn: Machine Learning in Python Fabian, Environ. Health Perspect. 127(9) (2019), 2825-2830, doi: 10.1289/EHP4713.

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002), 321-357 doi: 10.1613/jair.953.