International Journal of Advanced Information Science and Technology (IJAIST) ISSN: 2319:2682
Vol.3, No.10, October 2014 DOI:10.15693/ijaist/2014.v3i10.53-59

Application of Data Mining Methods in Employee Revenue Analysis

K. Tamizharasi[1]
[1]Research Scholar,
Periyar University,
Salem.

Dr. UmaRani[2]
[2]Associate Professor,
Sri Saradha College for
Women, Salem

K.Rajasekaran[3]
[3]Asst. Professor,
DB Jain College
Thoraipakkam, Chennai

Abstract:- The unique aspect of this research has been the use of five predictive data mining techniques on a sample data of 120 employees in an organization. The results of the study clearly show a relationship of employee turnover. The age and marital status emerged as key demographic variables. The findings of this study have implications for both research and practice. There is a need to expand the scope of this research to include multiple organizations and a large sample, which will allow for more robust predictions. For practitioners, it emphasises the need for greater use of models and analytical tools in engaging with human resource strategies and plans, and in particular that HR professionals will need to understand, appreciate and apply such models in future to be able to perform their roles as strategic business partners.

Index Terms :Data Mining, Employee Turnover, Applications, Algorithm

I.INTRODUCTION

Nowadays, in the K-Era, knowledge is a valued asset and among the crucial issues to address. Knowledge can be discovered through many approaches and one of them is by using data mining technique. In data mining, tasks such as classification, clustering and association is used to discover understood knowledge from the huge amount of data. Classification technique is a supervised learning technique in machine learning, which the class level or the target is already known. There are many fields adapted this approach as their problem solver method, such as finance, medical, marketing, stock market, telecommunication, manufacturing, health care, customer relationship, education and some others.

Nevertheless, the application of data mining has not attracted much attention in Human Resource Management (HRM) field (Chien & Chen, 2008; Ranjan, 2008). The vast amount of data in HRM can provide a rich resource for knowledge discovery and for decision support system development. Besides that, the valuable knowledge discovered from the data mining process should be considered as part of knowledge management issues. In any organization, they have to struggle effectively in term of cost, quality, service or innovation. The success of these tasks depends on having enough, right people with the right skills, employed in the appropriate locations at the appropriate point of time. This is categorized as part of the talent management task in HRM. In addition, talent management is a process to ensure the right person is in the right job (Cubbingham, 2007).

Recently, among the challenges of human resource professionals are managing an organization talent which involves a lot of managerial decisions. These types of decision are very uncertain and difficult. It depends on many factors like human experiences, knowledge, preferences and judgments. Besides that, the process to recognize the existing talent in an organization is among the top talent management issues and challenges (A TP Track Research Report 2005). Employees in an organization are assessed based on their performance in order to represent their talent ability. For that reason, this study aims to use classification techniques to classify the employee’s performance. In this case, the class level for the performance is whether the employee gets a recommendation for preferment or not. In this study, we use employee’s performance data from a selected organization as our data set. Therefore, the purpose of this paper is to suggest the possible classification techniques for employee future performance through some experiments using the selected classification algorithms. As a result, by using proposed classifier, we generate a prediction model which can be used for employee’s recital prediction. This paper is organized as follows. The second section describes the Data Mining Algorithms and Techniques. The third section discusses the related study. Section 4 shows methodology of the research. Finally, the paper ends at Section 5 with the concluding remarks and future research directions.

II.DATA MINING ALGORITHMS AND TECHNIQUES

Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases.

A. Classification

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and credit risk applications are particularly well suited to this type of analysis. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. For a fraud detection application, this would include complete records of both fraudulent and valued activities determined on a record-by-record basis.

The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier.

Types of classification models:

_ Classification by decision tree induction
_ Bayesian Classification
_ Neural Networks
_ Support Vector Machines (SVM)
_ Classification Based on Associations

III. RELATED WORK

A Naive Bayesian classifier is a simple probabilistic classifier based on applying Baye's theorem with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model"[2]. In simple terms, a Naive Baye's classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature, given the class variable.

Data mining tasks are generally categorized as clustering, association, classification and prediction (Chien & Chen, 2008; Ranjan, 2008). Over the years, data mining has evolved various techniques to perform the tasks that include database oriented techniques, statistic, machine learning, pattern recognition, neural network, rough set and etc. Database or data warehouse are rich with hidden information that can be used to provide intelligent decision making. Intelligent decision refers to the ability to make automated decision that is quite similar to human decision. Classification and prediction in machine learning are among the techniques that can produce intelligent decision. At this time, many classification and prediction techniques have been proposed by researchers in machine learning, pattern recognition and statistics.

Classification and prediction in data mining are two forms of data analysis that can be used to extract models to describe important data classes or to predict future data trends (Han & Kamber, 2006). The classification process has two phases; the first phase is learning process, the training data will be analyzed by the classification algorithm. The learned model or classifier shall be represented in the form of classification rules. Next, the second phase is a classification process where the test data are used to estimate the accuracy of the classification model or classifier. If the accuracy is considered acceptable, the rules can be applied to the classification of new data (Fig. 1). Several techniques that are used for data classification are decision tree, Bayesian methods, Bayesian network, rule-based algorithms, neural network, support vector machine association rule mining, k-nearest-neighbor, case-based reasoning, genetic algorithms, rough sets, and fuzzy logic. In this study, we attempt to use three main classification techniques, i.e. decision tree, neural network and k-nearest-neighbor. However, decision tree and neural network are found useful in developing predictive models in many fields(Tso & Yau, 2007). The advantage of the decision tree technique is that it does not require any domain knowledge or parameter setting, and is appropriate for exploratory knowledge discovery. The second technique is a neural-network, which has high tolerance of noisy data as well as the ability to classify patterns on which they have not been trained. It can be used when we have little knowledge of the relationship between attributes and classes. Next, the K-nearest-neighbor technique is an instance-based learning using distance metric to measure the similarity of instances. All these three classification techniques have their own advantages and disadvantages, for that reason, this study endeavor to explore these classification techniques for human talent data. Besides that, data mining technique has been applied in many fields, but its application in HR is very rare (Chien & Chen, 2008).

IV.METHODOLOGY

A. Study Site and Respondents

The sample consisted of employees who had left the company during the past three years as well as those who are still with the company at the time of selecting the sample. Among all the employees in the sample, 15 per cent had left the company at the date of sample selection. The sample was predominantly male (85 percent). Only one third of the sample was married, and the respondents were relatively young with only 32 per cent aged above 30 years. The average total work experience was slightly less than five years, with an average experience within the company slightly more than two years. The average experience within the current team was slightly more than 15 months, which is somewhat longer than the usual norm in the industry. The average time in the current position was less than 15 months. This condition indicates that promotions were both rapid and prompt, despite this employment being the first job for about one third of the respondents, and almost all employees had changed their job position once. In summary, it appears that the employees in the sample are young, employed in a rapidly growing company and continuing in the same team for a reasonably long time.

B. Procedure

The study was lead using secondary data available in the archives of the organisation. The data on employee turnover were obtained from an industry and a sample of 120 employees was extracted from the database of the company. To guarantee the confidentiality of information, the names and other personal details of the employees were removed from the data. In order to facilitate validation of the models, each employee record was given a unique identification number.

All the experience related variables were categorized. It was based on equi-depth binning. The age of the employee was derived from the date of birth in the company database. The month wise data on casual leave, medical leave, and daily arrival times were analyzed to identify changes in the patterns during the past six months. The analysis was primarily aimed at isolating cases where the usage was similar, or increasing or decreasing over the past six months. When an employee had left the company, the data for six months prior to leaving the company was analysed. A similar analysis was done to identify changes in patterns in arrival times at work. The normal practice while applying data mining techniques is to divide the data into training and testing data sets. Such division is usually done on a random basis. The models are trained using the training data set and then the model thus, developed is tested using the testing dataset. The main objective of such separation of training and testing datasets is to make sure that the models developed will not be specific to the special patterns in a particular dataset. Such a separation would be possible where the number of observations is large enough to allow such a luxury. In the present case the same dataset was used for training as well as for testing because the dataset contained only a limited number of observations.

C. Measures

The following data were obtained from the employee records:

- Date of birth
- Gender
- Marital status
- Total years of work experience (in five categories)
- Months of experience in the present company (in five categories)
- Months of experience in the current team (in five categories)
- Months of experience in the current position (in five categories)
- Type of position occupied currently in the company (in six categories)
- Type of domain expertise
- Frequency of job change till joining the present company (in five categories)
- Month wise use of casual leave (in five categories)
- Month wise use of medical leave (in five categories)
- Month wise data on arrival time at work (in five categories)

D Analysis

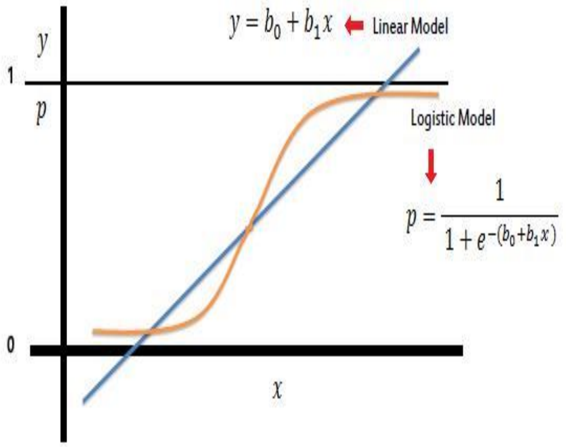
Five different prediction models were used and data were trained and tested on them. The methodology adopted application of various data mining techniques to predict employee turnover. The models used are artificial neural networks, logistic regression, classification and regression trees, classification trees (C5.0), and discriminant analysis.

E. Logistic Regression

Logistic regression predicts the probability of an outcome that can only have two values (i.e. a dichotomy). The prediction is based on the use of one or several predictors (numerical and categorical). A linear regression is not appropriate for predicting the value of a binary variable for two reasons:

A linear regression will predict values outside the acceptable range (e.g. Predicting probabilities outside the range 0 to 1). Since the dichotomous experiments can only have one of two possible values for each experiment, the residuals will not be normally distributed about the predicted line.

On the other hand, a logistic regression produces a logistic curve, which is limited to values between 0 and 1. Logistic regression is similar to a linear regression, but the curve is constructed using the natural logarithm of the "odds" of the target variable, rather than the probability. Moreover, the predictors do not have to be normally distributed or have equal variance in each group.



In the logistic regression the constant (b0) moves the curve left and right and the slope (b1) defines the steepness of the curve. By simple transformation, the logistic regression equation can be written in terms of an odds ratio.

$$\frac{p}{1-p} = \exp\left(b_0 + b_1 x\right)$$

Finally, taking the natural log of both sides, we can write the equation in terms of log-odds (logit) which is a linear function of the predictors. The coefficient (b1) is the amount the logit (log-odds) changes with a one unit change in x.

$$ln\left(\frac{p}{1-p}\right) = b_0 + b_1 x$$

As mentioned before, logistic regression can handle any number of numerical and/or categorical variables.

$$p = \frac{1}{1+e^{-(b_0+b_1x_1+b_2x_2+\cdots+b_px_p)}}$$

There are several analogies between linear regression and logistic regression. Just as ordinary least SQUARE regression is the method used to estimate coefficients for the best fit line in linear regression, logistic regression uses maximum likelihood estimation (MLE) to obtain the model coefficients that relate predictors to the target. After this initial function is estimated, the process is repeated until LL (Log Likelihood) does not change significantly.

$$\beta^1 = \beta^0 + [X^T W X]^{-1}.X^T(y-\mu)$$

β is a vector of the logistic regression coefficients.

W is a square matrix of order N with elements $n_i\pi_i(1-\pi_i)$ on the diagonal and zeros everywhere else.

μ is a vector of length N with elements $\mu_i = n_i\pi_i$.

A pseudo R2 value is also available to indicate the adequacy of the regression model. Likelihood ratio test is a test of the significance of the difference between the likelihood ratio for the baseline model minus the likelihood ratio for a reduced model. This difference is called "model chi-square“. Wald test is used to test the statistical significance of each coefficient (b) in the model (i.e., predictors contribution).

F. Classification Trees (C5.0)

In the case of C5.0 classification trees, the splitting of the records at each node is done based on the information gain. Entropy is used to measure the information gain at each node. This method can generate trees with variable number of branches at each node. For example, when a discrete variable is selected as an attribute for splitting, there would be one branch for each value of the attribute. The construction of the tree, creation of leaf nodes and labelling of the leaf nodes as well as the estimation of error rates are very similar to the CART methodology.

G. Discriminant Analysis

Discriminant analysis is one of the commonly used statistical techniques where the dependent variable is categorical or nominal in nature and the independent variables are metric or ratio variables. It involves deriving a variant or z-score, which is a linear combination of two or more independent variables that will discriminate best between two (or more) different categories or groups. The discriminant analysis involves creating one or more discriminant function so as to maximize the variance between the categories relative to the variance with the categories. The z-scores calculated using the discriminant functions could be used to estimate the probabilities that a particular member or observation belongs to a particular category. It is important that the independent variables used in discriminant analysis are continuous or metric in nature. Accordingly, the variables used in estimating the discriminant function are the original variables.

V. RESEARCH FINDINGS

Application of basic statistical methods is used to study the employee. It is found that women turnover comparatively less than men. To carry out a more detailed turnover analysis and separate the attributes which have the highest effect, decision trees are used.

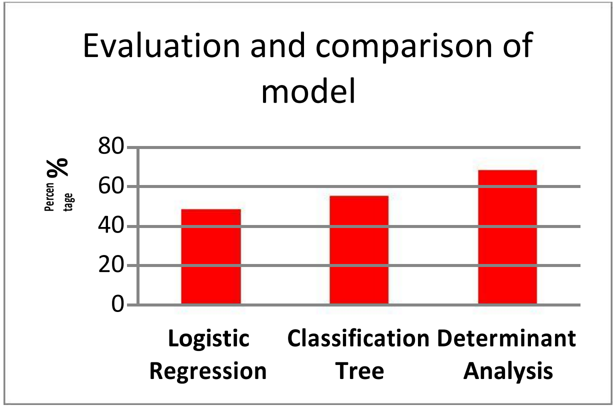


Figure 1.1 : Evaluation and Comparison of model

The Figure 4 represents the basis or the 40% of employees from the original population that turnover. The graph shows that the model Neural-2 is evaluated as the best one in comparison to all the other models (it is positioned upwards). Other models are more difficult to interpret due to

their overlapping. Figure 4 also shows that neural networks provide better results when using the missing values replacement method "Tree imputation" rather than "Most frequent value". The graph can be read in the following way: the plot Neural-2 shows that there are approximately 60% of employees that will turn over.

VI. CONCLUSION

Three salient conclusions can be made. First, the study establishes the value of the use of prediction models to identify and predict voluntary employee turnover in organisations. While the overall predictive accuracy was very high across all models, in the current study it appears that the best prediction was possible with discriminant analysis. Secondly, the identification of the five variables, namely demographics, tenure, job content and employee turnover in the discriminant analysis is significant from a research perspective. Thirdly, while the predictive accuracies are specific to the data used in the analysis and to the specific company studied, the study has shown that it is possible to predict the employee turnover, and identify those who have turnover intentions even before they had made their final decision to leave.

This study raises several issues for future research. First, further research could explicitly collect data on demographic variables across a large sample of organisations to examine the relationship between demographic variables and turnover. Second, large scale data on variables in the past academic research which have a relationship with turnover can be collected longitudinally. Such a data set will allow for more rigorous analysis and also a refined prediction model. Third, the context specific variables of employee turnover, which emerged from this study would warrant a deeper understanding of the phenomena. There is a need for more empirical research and in particular, longitudinal research using data within corporations to refine the model. Last, more research needs to be conducted in various different samples to confirm the validation of the theoretical model and the prediction model proposed in this study.

This research has implications for HR professionals and practising managers. There is a growing recognition that human resources are the source of competitive advantage for organisations in a global economy. Knowledge and services, the two key sectors of the modern economy are people centred and people driven businesses. Therefore, tools and models that enhance understanding and prediction of any attitudinal and behavioural variables can bring significant value to practitioners. In recent years, various authors have urged human resource professionals to play the role of a strategic partner (Ulrich & Brockbank 2005). Usage of these prediction models with the existing organisational data is likely to enhance the image and effectiveness of the HR professionals and departments. The use of such models to predict turnover allows firms to formulate targeted retention strategies with an aim to ensure that key people stay with the organisation and that wasteful and expensive levels of employee turnover are reduced. The prediction models present managers reliable and accurate information on the antecedents and factors that cause people to leave. It also provides an opportunity for managers to make more data driven decision making in organisation on people related issues.

REFERENCE

[1] Black J &Gregersen B 1999, ‘The right way to manage expats’, Harward Business Review, vol. 77, no. 2, pp. 52-63.

[2] Carey, D. C., & Ogden, D. (2004). The human side of M&A: Leveraging the most important factor in deal making. New York: Oxford University Press.

[3] Cf. A. Krasnodębska, Occupational preferences of Opole students and the issue of labour-related migration abroad,WydawnictwoInstytutŚląski, Opole 2008.

[4]Gamberger, D., Šmuc, T. (2001): Poslužiteljzaanalizupodataka (http://dms.irb.hr). Zagreb, Hrvatska:InstitutRudjerBošković, Laboratorijzainformacijskesustave.

[5] Good communication can help boost both your bottom line and employee retention. (2004, September). Contractor’s Business Management Report, 5-7.

[6] Hair, J., Anderson, R., Babin, B. (2009): Multivariate Data Analysis, Prentice Hall.

[7] Halmi A. (2003): Multivarijantnaanaliza u društvenimznanostima, Alinea, Zagreb

[8] Hurn, J 1999, ‘Repatriation - the toughest assignment of all’, Industrial and Commercial Training, vol. 31, no. 6, pp. 224-228.

[9] J. Guichard, M. Huteau, Counselling Psychology, OficynaWydawnicza “Impuls”, Krakow 2005, p.14.

[10] Matignon, R. (2007): Data Mining Using SAS Enterprise Miner TM, John Wiley & Sons, Inc., Hoboken, New Jersey

[11] Matignon, R. (2007): Data Mining Using SAS Enterprise Miner TM, John Wiley & Sons, Inc., Hoboken, New Jersey

[12] Olson, D.L., Delen, D. (2008): Advanced Data Mining Techniques, Springer-Verlag, Berlin Heidelberg.

[13] Paik, Y, Segaud B & Malinowski C 2002, ‘How to improve repatriation management’, International Journal of Manpower, Vol. 23, no 7, pp. 635-675.

[14] Panian, Ž., Klepac, G. (2003): Poslovnainteligencija, Masmedia

[15] SAS Institute (2004): Getting Started with SAS Enterprise Miner 4.3, Second Edition, SAS Institute Inc., Cary

[16] Stroh, L 1995, ‘Predicting turnover among repatriates: can organizations affect retention rates?’, The International Journal of Human Resource Management, vol. 6, no.2, pp. 443-456.

[17] Stroh, L, Gregersen, B & Black, J 1998, ‘Closing the Gap: Expectations Versus Reality Among Repatriates’, Journal of World Business, vol. 33, no. 2, pp. 111-124

[18] Tanguy, 1986, see: ibidem, p. 21.

[19] Thornton, S. L. (2001, October/November). How communication can aid retention: Making communication a management priority. Strategic Communication Management, 24-27.

[20] Ulschak, F.L., &Snowantle, S.M. (1992). Managing employee turnover; A guide for health care executives. Chicago, Illinois: American Hospital Publishing.

[21]Westcott, S. (2006, April). Goodbye and good luck. Inc. Magazine, 40-42.

[22] Bretz, R. D., Boudreau, J. W., & Judge, T. A. (1994). Job search behaviour of employed managers. Personnel Psychology, 47(2), 275-301.

[23] Chaudhuri, K. K. (2007). Managing 21st C employees. Personnel Today, (Jan-March) 18-19.

[24] Crosby, J. V., & Brandt, D. M. (1988). Age and voluntary turnover: A quantitative review. Personnel Psychology, 48(2), 335-345.

[25] Hammer, T. H., Landau, J., & Stern, R. N. (1981) Absenteeism when workers have a voice: The case of employee ownership. Journal of Applied Psychology, 66(5), 561-573.

[26] Johns, G. (1995). Absenteeism. In N. Nicholson (Ed.), The Blackwell Encyclopaedic Dictionary of Organizational Behaviour (1-3). Oxford, UK: Blackwell.

[27] Lee, T. W., & Mitchell, T. R (1994). An alternative approach: The unfolding model of voluntary employee turnover. Academy of Management Review, 19(1), 51-89.

[28] Louis, M. (1980). Surprise and sense making: What newcomers experience in entering unfamiliar organizational settings. Administrative Science Quarterly, 25(2), 226-251.

[29] Barney, J. B., & Wright, P. M. (1998). On becoming a strategic partner: The role of human resources in gaining competitive advantage. Human Resource Management, 37(1), 31-46

Author Profile

K. Tamizharasi Working as Asst.Professor in the Department of Computer Application,SIVET, Chennai Tamilnadu. She has 6 years of experience in Industry and academic fields. She completed her MCA in Bharathiar University, Coimbatore and M.Phil from Vinayaka Missions University, Salem. Now, doing as Research Scholar in Periyar University, Salem in the field of Computer Science. Her area of Interest includes Data Mining. She has also life member for several association and society.

Dr. R.Uma Rani received her Ph.D., Degree from Periyar University, Salem, Tamil Nadu, India in the year 2006. She is a rank holder in M.C.A., from NIT, Trichy. She has published around 40 papers in reputed journals, national and international conferences. She has received the best paper award from VIT,Vellore, Tamil Nadu in an international conference. She has done one MRP funded by UGC. She has acted as resource person in various national and international conferences. Her areas of interest include Information Security, Data Mining, Fuzzy Logic and Mobile Computing.

K. Rajasekaran Working as Assistant Professor in DB Jain College, Thoraipakkam, Chennai Tamilnadu. He has 5+ years of experience in Industry and academicfields. He completed his M.Phil(Computer Science) in Periyar University, Salem. And MCA in Madras University, Chennai. He has also life member for several association and society.