ESTABLISHING A LINK BETWEEN EMPLOYEE TURNOVER AND WITHDRAWAL BEHAVIOURS: APPLICATION OF DATA MINING TECHNIQUES

Abstract

Employee turnover is a serious concern in knowledge based organisations. When employees leave an organisation, they carry with them invaluable tacit knowledge which is often the source of competitive advantage for the business. In a rapidly growing sector like the Indian software industry employee turnover poses risk and challenges for organisations. This research explores the relationship of withdrawal behaviours like lateness and absenteeism, job content, tenure and demographics on employee turnover. The unique aspect of this research has been the use of five predictive data mining techniques on a sample data of 150 employees in a large software organisation. The results of the study clearly show a relationship between withdrawal behaviours and employee turnover. 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 organisations 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.

Submitted
20 October 2007
Resubmitted
3 April 2008
Resubmitted
30 August 2008
Final version
1 October 2008
Refereed anonymously

Key Words: India, Software Industry, Employee Turnover, Withdrawal Behaviour, Data Mining Techniques

Vishnuprasad Nagadevara
Indian Institute of Management Bangalore
India

Vasanthi Srinivasan
Indian Institute of Management Bangalore
India

Reimara Valk
Indian Institute of Management Bangalore
India

INTRODUCTION

In the knowledge economy, human resources with knowledge and competencies are the key assets in assisting firms and/or countries to sustain their competitive advantage. Globally competitive organisations will depend on the uniqueness of their human resources and the systems for managing human resources effectively to gain competitive advantage (Pfeffer 1994, Bartlett & Ghoshal 1997, Barney & Wright 1998). Human resources are not only the drivers and principal value creators of the output of the knowledge industry, but they are also the intellectual capital or the 'infrastructure investment'. Therefore, attracting, training, retaining and motivating employees are the critical success determinants for any knowledge based organisation.

The phenomenal growth of the Indian software industry is well documented (Arora & Athreye 2002, Rajawat 2007). It is estimated that the industry would employ 850,000 IT professionals and 1.4 million BPO personnel by 2010. Existence of English speaking technically qualified manpower, competitive billing, high productivity gains and scalability are the reasons attributed for the emergence of India as a key IT services outsourcing destination (*Nasscom* 2006). As the demand for software professionals increased with supply increasing marginally, wages in the software industry began to spiral. While the domestic demand increased because of the growth, the overseas demand for Indian software engineers rose out of significant IT skills shortage across the world. All of this has meant an increase in the turnover rates in the industry. A recent survey estimated the turnover rate in software firms in India as high as 25 to 60 per cent (Ramani & Raghunandan 2008). When IT professionals leave an organisation, not only are less of them available for assignment to project, but the professionals themselves often take specialised skills, tacit knowledge, and understanding of specific business operations and information systems with them (Moore & Burke 2002). In addition, when turnover is at senior levels it is likely that a firm loses some of its key clients as they move with the employee. In recent years, the ability of the firms to understand and anticipate voluntary turnover is becoming a key requirement.

Turnover results can have direct and indirect, tangible and intangible costs and a loss of social capital, which may impact organisational success (Dess & Shaw 2001). The tangible costs would be the cost of training new employees, the recruitment and selection costs, adjustment time, possible product and/or service quality problems, costs of agency workers/ temporary staff (Morrell, Loan-Clarke & Wilkinson 2004), the cost of loss productivity, the cost of lost knowledge and the cost of the position remaining vacant till a suitable replacement is found (Sharma 2007). For knowledge intensive activities such as high tech product development, turnover means not only losing people to competitors, but also knowledge walking out of the organisation (Moitra 2001). Given the phenomenal growth of the Indian software industry over the last two decades, it is important for organisations to understand, predict and control employee turnover. The use of statistical models to understand these phenomena, therefore, becomes relevant in the current context. Application of data mining techniques to behavioural issues has been established by other authors (Somers 1999, Nagadevara 2004, 2008)

The purpose of this paper is to evaluate a conceptual model that is believed to predict employee turnover in the software industry in India using data mining techniques. Indeed, the results of this investigation may provide a sounder underpinning for understanding the processes that can lead to severance of individual-organisational linkages. Five predictive models, which have been used to understand the phenomenon of voluntary turnover, are employed. The use of these models is likely to significantly contribute to better understand the phenomenon of turnover, and thus, enhance the efficiency and effectiveness of those human resource planning processes that are used to focus on this persistent problem.

VOLUNTARY TURNOVER OF EMPLOYEES

Employee turnover is a widely researched phenomenon. There have been a large number of studies (Shaw, Delery, Jenkins & Gupta 1998) and meta analysis on this subject (Morrell, Loan Clarke & Wilkinson 2004). However, there is no universally accepted account for why people choose to leave (Lee & Mitchell 1994). Booth and Hamer (2007) found that labour turnover is related to a variety of environmental factors and organisational factors such as company culture and values, supervisory style, fair pay, corporate value, giving support to each other, trust and respect between employees, manageable workload, development and career building satisfaction and degree of job satisfaction. Employee turnover is a function of many different psychological states, including job dissatisfaction (Mobley 1977, Tett & Meyer 1993), lack of organisational commitment (Porter, Crampion & Smith 1976, Tett & Meyer 1993) and availability of alternative jobs (McLaughlin 1991, Bretz, Boudreau & Judge 1994). Research conducted by Chaudhuri (2007) shows that the causes of turnover in the Indian software industry are unchallenging work environments, long working hours, limited career growth, less promotional opportunities, lack of proper leadership, non attractive compensation packages, job opportunities elsewhere and poaching of talent by the competitors. Indeed, voluntary turnover, popularly termed ‘job hopping’, has been a persistent problem for human resource management practitioners.

Labour market variables (Kirschenbaum & Mano-Negrin 1999), organisational and occupational employment opportunities and occupational preferences (Mano-Negrin 2001) and equity (Aquino, Griffeth, Allen & Hom 1997) are known to result in voluntary turnover. Withdrawal behaviours like lateness and absenteeism are known to be precursor behaviours to voluntary turnover.

Withdrawal Behaviours: Lateness and Absenteeism

The three dimensions namely turnover, absenteeism and lateness have been grouped together as withdrawal behaviours and have been studied by work psychologists. Turnover is "…the termination of an individual's formal membership with an organisation." (Lee 1997:97). Various authors have defined absenteeism as the "…failure to report for scheduled work." (Johns 1995:1) or "…an individual's lack of physical presence at a given location and time when there is a social expectation for him or her to be there." (Martocchio & Harrison 1993: 259). Lateness is defined as the "…the tendency of an employee to arrive at work after the scheduled starting time." (Adler & Golan 1981:544). The linkages between the three kinds of withdrawal behaviours have been of interest to researchers. In particular, the progression perspective that withdrawal will progress from minor, less salient acts, such as occasional lateness, to more salient acts, such as absence and finally turnover (Johns 1995) has gained importance. Longitudinal studies by Wolphin, Burke, Krausz and Freibach (1988) and Rosse (1988) found a lateness-absence progression although Adler and Golan (1981) and Krausz, Koslowsky and Eiser (1998) did not. Blau (1994) found a pattern of increasing chronic lateness that was associated with elevated absence within the same 18 month period. Several studies reveal a progression from absence to turnover (Crosby & Brandt 1988, Krausz, et al. 1998, Rosse 1988). If the progression was indeed existent, then statistically it would expect the relationship between lateness and absenteeism, absenteeism and turnover to be stronger than the relationship between lateness and turnover. The meta analysis studies support this connection. Kozlowsky and Salas (1997) reported a corrected correlation of 0.40 between lateness and absence and Mitra, et al (1992) reported a corrected correlation of 0.33 between absence and turnover. Kozlowsky and Salas (1997) estimated the mean correlation between lateness and actual turnover to be 0.07 and that between lateness and an apparent composite of actual turnover and turnover intentions to be 0.27 (Johns 1995).

Demographic Variables: Age, Gender and Marital Status

Personal demographic variables have been widely touted as absence predictors. For instance, studies have demonstrated that gender and age (Clegg 1983) and education level (Hammer, Landau & Stern 1981) all have some validity as predictors of absenteeism. The literature on the impact of demographic variables like age, tenure, gender and education on withdrawal behaviours has been ambiguous. Some studies have reported a positive relationship between absenteeism and educational level and gender (Steel & Rentsch 1995). In a review of literature on turnover among sales professionals, Lucas, Parasuraman, Davis and Enis (1987) report that, among all personal characteristics, the most studied and the most consistent in its relationship to turnover is the employee's age. Older employees were less likely to leave the organisation than younger employees. Some studies show through quantitative evidence that age alone is an insufficient predictor of voluntary turnover. In other words, age provides little information for judging when an employee will choose to leave an organisation. In their research on the IT industry in India, Ahuja, Chudoba, Kacmer, Mcknight and George (2007) found age had a modest but significant effect on turnover intention. Gender and marital status also did not affect turnover intention. However, marital status would become an important variable in the context of the Indian software industry. A number of employees are young in age and many employees voluntarily turnover when their spouses change jobs or relocate to other cities. Clearly, there is merit in examining the relationship between demographic variables like age, gender, marital status and turnover in the context of the Indian software industry.

Tenure

The relationship between tenure and turnover in organisations is not clearly established. However, some researchers (Mobley 1977) consider tenure to be the best single predictor of turnover. Turnover amongst newcomers is often caused by unmet expectations and violation of the psychological contract. Expectations held by newcomers are often inflated or unrealistic (Wanous 1992). This could be because people often compare their actual job experiences with early job experiences in other organisations (Louis 1980). Hence, prior work experience seems to influence employee turnover in such a way that if early job experiences do not meet expectations turnover is likely. This phenomenon is also supported by other empirical research on turnover in the IT industry in India (Rathi 2003).

Job Content

The software services industry in India is characterised by the execution of large scale software projects for global clients. These software projects such as product development or consulting require understanding of the client expectations. Traditionally Indian companies have been engaged largely in the lower end operations such as coding and providing standardised solutions. During the last decade, many large Indian organisations have moved up the value chain (*Nasscom* 2006) by providing higher value services, and such high value services require employees to demonstrate and work on cutting edge technologies and state of the art domain knowledge. The nature of the work and role played by the employee in the organisation has an impact on turnover. In an empirical study of employee turnover in the IT industry in India, Rathi (2003) identified the domain of work as one of the factors for employees leaving the organisation. Thus, organisations which provide opportunities on cutting edge technologies and state of the art domain knowledge are likely to experience lower rates of turnover.

Frequency of Job Change

In industry contexts, which are experiencing high rates of growth, the demand and supply of skilled labour gets skewed. In the case of the software industry in India, the talent shortage arising out of the demand and supply situation is significant as most young and competent engineers have multiple job offers. The maturity exercised by them in the job choice determines their longevity in the organisation. This element is captured in the tenure literature. However, it is not just the tenure, but the frequency with which they turnover is also important as it is likely that at the exploratory stages of career choice and decision making, many young adults would focus on crystallising an occupation choice and then identify the organisations in which they would like to be employed. In general, people tend to choose organisations that they view as most instrumental in helping them attain their work preferences. Studies of the recruitment process (with student populations) have shown that factors such as familiarity with a specific employer and perceived employer attributes like company image, rewards and job challenges are positively related to applicant attraction (Hall 2002). After the candidate joins the organisation, the ‘reality shock’ of the actual work experience impacts their attitude to the organisation, and in the context of rapid growth where large numbers of employees are being hired, it is likely that the recruiters may paint a rosier picture of the job than what it actually is. The career exploration process combined with the ‘reality shock’ could potentially result in a greater frequency of job changes.

The objective of the study was to understand and predict the role of demographics, tenure, job content, frequency of job change and withdrawal behaviours on employee turnover using data mining techniques. The theoretical model describing the relationship across these variables is given in Figure 1. This conceptual model has been generated from the presented theoretical contentions.

(Insert Figure 1 here)

Figure 1 posits employee turnover is influenced by numerous contextual and personal attributes. In practice these variables can be behavioural features of the work context (e.g., tenure), the job content dimensions as well as the demographics of the job holder. The variable of tenure is measured by total work experience, experience in current team and current position. One of the good predictor variables for turnover is withdrawal behaviour, which is indicated by absenteeism and lateness. These withdrawal behaviours are in turn influenced by the type job content and tenure. The job content involves the type of position that the employee occupies in the organisation and the type of domain expertise. The important demographic characteristics are age, gender and marital status. This conceptual model is used to establish the linkage between turnover and withdrawal behaviour and these linkages, are the hypotheses that are depicted as arrow headed lines.

METHODOLOGY

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, 28 per cent had left the company at the date of sample selection. The sample was predominantly male (70 per cent). Only one third of the sample was married, and the respondents were relatively young with only 30 per cent aged above 28 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 18 months, which is somewhat longer than the usual norm in the industry. The average time in the current position was less than 18 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.

Procedure

The study was conducted using secondary data available in the archives of the organisation. The data on employee turnover were obtained from a software company and a sample of 150 employees was extracted from the records 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 categorised. It was based on equi depth binning. The age of the employee was derived from the date of birth. The month wise data on casual leave, privilege leave and daily arrival times were analysed 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. 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 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.

Measures

The following data were obtained from the employee records:

- Date of birth
- Gender
- Marital status
- Total years of work experience (in three categories)
- Months of experience in the present company (in three categories)
- Months of experience in the current team (in three categories)
- Months of experience in the current position (in three categories)
- Type of position occupied currently in the company (in six categories)
- Type of software domain expertise

- Frequency of job change till joining the present company (in three categories)
- Month wise use of casual leave (in three categories)
- Month wise use of privilege leave (in three categories)
- Month wise data on arrival time at work (in three categories)

Analysis

Five different prediction models were used and data was 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.

Artificial Neural Networks (ANN)

The ANNs are generally based on the concepts of the human (or biological) neural network consisting of neurons, which are interconnected by the processing elements. The ANNs are composed of two main structures namely the nodes and the links. The nodes correspond to the neurons and the links correspond to the links between neurons. The ANN accepts the values of inputs into what are called input nodes. This set of nodes is also referred to as the input layer. These input values are then multiplied by a set of numbers (also called as weights) that are stored in the links. These values, after multiplication, are added together to become inputs to the set of nodes that are to the right of the input nodes. This layer of nodes is usually referred to as the hidden layer. Many ANNs contain multiple hidden layers, each feeding into the next layer. Finally, the values from last hidden layer are fed into an output node, where a special mapping or thresholding function is applied and the resulting number is mapped to the prediction.

The ANN is created by presenting the network with inputs from many records whose outcome is already known. For example, the data on age, income and occupation of the first employee (first record) are inputted into the input layer. These values are fed into the hidden layer and after processing (by combining these values using appropriate weights) the prediction is made at the output layer. If the prediction made by the ANN matches with the actual known status of the employee (say either left the company or not), then the prediction is good and the ANN proceeds to the next record. If the prediction is wrong, then the extent of error (expressed in numerical values) is apportioned back into the links and the hidden nodes. In other words, the values of the weights at each link of each node in the hidden layers are modified based on the extent of error in prediction. This process is referred to as the backward propagation. The artificial neural networks are found to be effective in detecting unknown relationships. ANNs have been applied in many service industries such as health (to identify the length of stay and hospital expenses) (Nagadevara 2004), air lines and hospitality (Chatfield 1998, Nagadevara 2008).

A total of 14 variables are used to build the ANNs for predicting the employee turnover. These are in addition to the dependent variable, which is a nominal variable indicating whether the employee is still with the company or had left the company. Most often the mathematical relationships or equations developed by the ANNs are complex and not available to the user. As a result, these are treated as black boxes, only to be used to obtain the prediction results. Nevertheless, it is important to know the relative importance of each of the variables in predicting the categories. The software used to build the ANNs provides the sensitivity of the prediction with respect to each of the variables and this can be viewed as an indicator of the relative importance of the variables in question.

Logistic Regression

Logistic regression is a specialised form of regression used to predict and explain a categorical dependent variable. It works best when the dependent variable is a binary categorical variable. The regression equation developed is very similar to a multiple regression equation with regression like coefficients which explain the impact of each of the independent variables in predicting the category of the dependent variable. One special advantage of logistic regression is that it is not restricted by the normality assumption which is a basic assumption in the regression analysis. This technique can also accommodate non metric variables such as nominal or categorical variables by coding them into dummy variables. Another advantage of logistic regression is that it directly predicts the probability of an event occurring. In order to make sure that the dependent variable, which is the probability, is bounded between zero and one, the logistic regression defines a relationship between the dependent and independent variables that resembles an S-shaped curve, which uses an iterative process to estimate the ‘most likely’values of the coefficients. This results in the use of a ‘likelihood’ function in fitting the equation rather than using the sum of squares approach of the regression analysis. The dependent variable is considered as the ‘odds ratio’ of a specific observation belonging to a particular group or category. In that sense, logistic regression estimates the probability directly.

In order to get the best prediction results from the logistic regression, it is important to have continuous variables as independent variables. It is also important to define the nominal variables appropriately, so that they are converted into the required number of dummy variables. Thus, the use of categorised variables is kept to the minimum possible in the case of logistic regression.

Classification and Regression Trees (CART)

Classification and Regression Trees (CART) is one of the popular methods of building classification trees. CART always builds a binary tree by splitting the observations at each node based on a single attribute or variable. CART uses a measure called Gini Index for identifying the best split. If no split that could significantly reduce the diversity of a given node could be found, the process of splitting is stopped and the node is labelled as a leaf node. When all the nodes become leaf nodes, the tree is fully grown. At the end of the construction of the tree, each and every observation has been assigned to a leaf node. Each leaf can now be assigned to a particular class and a corresponding error rate. The error rate at the leaf node is nothing but the percentage of misclassifications at the leaf node. The error rate for the entire tree is the weighted sum of the error rates of all the leaf nodes.

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.

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 variate 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 functions so as to maximise 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.

RESULTS

Table 1 summarises the information with respect to each of the 14 variables used for building the ANNs. The weights that are used in building the ANNs are not available to the user. On the other hand, the software used to build the ANNs carries out a sensitivity analysis by perturbing the values of each of the variables and measuring the resultant impact on the accuracy of the prediction. Based on this sensitivity analysis, the relative importance of each of the variables used in the model is calculated and presented in Table 1. Two of the five most important variables happen to be behavioural variables namely privilege leave used and late arrival at the office.

[Insert Table 1 here]

The variables used in building the logistic regression and the corresponding coefficients are presented in Table 2. Also Table 2 presents not only the regression coefficients, but also its exponential value. The relative importance of different variables on the 'odds ratio' can be obtained directly from Table 2. The chi-square distribution is used to test the statistical significance of the logistic regression model. The calculated value of chi-square for the model is 60.179 with 16 degrees of freedom, indicating a very high significance level. The nominal variables and quantitative variables used in the logistic regression are presented separately in Table 2. The behavioural variable namely casual leave used is one of the most important variables for predicting turnover.

[Insert Table 2 about here]

The classification and regression trees work best with nominal or binned variables. Hence, the data used to build the classification and regression tree is either in the form of binned data or nominal data. The resultant tree is presented in Figure 2. It can be seen from the tree that experience in the current team is one of the important determinants. Not only it appears at the top of the tree, but it also results in a pure node on one of the branches. Similarly, the relative importance of different variables could be seen from the tree based on their location in the tree.

[Insert Figure 2 here]

Classification Trees (C5.0) also are best suited for nominal or binned variables. The tree constructed using the classification tree (C5.0) method is presented in Figure 3. One major difference between the CART and the classification trees is that classification trees allow multiple branches at any given node where as CART allows only binary splits. This is evident at Node 4 (Level 3) of the classification tree. In the case of classification trees also, experience in the current team appears to be one of the important factors in predicting turnover. The pattern of late arrival at the office appears to be one of the important determinants under classification trees. The prediction accuracy of the classification trees is 88.44 per cent, which is very similar to that of the CART. The prediction accuracy of the classification tree is marginally better with respect to those who had left the company (as compared to CART). The prediction accuracies of both CART and Classification Trees are very similar. This is not unexpected, considering that both the techniques work on similar principles, but differ only in terms of methodologies and splitting criteria adopted.

[Insert Figure 3 here]

Another classification technique used for predicting turnover is the discriminant analysis. One of the commonly used tests of statistical significance for measuring the effectiveness of the discriminant function is Wilks' Lambda. The smaller the value of Wilks' Lambda, the better is the discriminating ability of the function. In this particular case the value of Wilks Lambda is 0.742. The statistical significance of Wilks' Lambda is tested by converting it into an approximation of the chi-square distribution. The calculated value of the chi-square in this particular case is 37.873 with 10 degrees of freedom, indicating a very high level of statistical significance.

The prediction accuracy of employee turnover based on discriminant analysis was 82.09 per cent. The predictions are more accurate with respect to the employees who had left the organisation. The coefficients of the standardised discriminant function are presented in Table 3. It can be seen from Table 3 that the standardised coefficients with respect to four variables namely, age, experience in the company, casual leave used and late arrival at the office are negative. The dependent variable, which is categorical, is coded as '0' for those who left the company and as ‘1’ for those who remained in the company. Thus it can be concluded that these variables with a negative coefficients have a positive relationship with turnover.

[Insert Table 3 here]

As shown in Table 4, the prediction accuracy obtained using the ANNs is 81.63 per cent. While the ANNs are able to predict those who remained with the company with an accuracy level of more than 90 per cent, the accuracy level of prediction is only 59 per cent with respect to those who left the company. In other words, the ANNs are excellent in their predictions with respect to those who remain with the company. The most important indicator for prediction is the pattern of use of Privilege leave. This is followed by the current position in the company, current technical expertise and the pattern of late arrival at the office. It is not surprising that two of the four most important indicators are the changes in the behavioural patterns of the employees, proving that certain withdrawal behaviours are strong indicators of turnover.

[Insert Table 4 here]

Table 4 also shows the prediction accuracy obtained by the logistic regression is 79.58 per cent. In logistic regression, interpretation of the regression coefficient is not the same as that of the regular regression equation. The exponential value of the coefficient is considered to be the measure of the impact of the corresponding independent variable on the ‘odds-ratio’. The results indicate that marital status has an impact on turnover. Married people are less likely to turnover than singles.

In addition, Table 4 shows the prediction accuracy of the Classification and Regression Tree is 89.80 per cent. The predictions are more robust with respect to those who have not left the company. Similarly, predictions of the Classification Trees (C5.0) are more accurate with respect to those who have not left the company. Overall, the prediction accuracies of the Classification and Regression Trees and Classification Trees (C5.0) are very similar to each other.

Table 4 summarises the prediction accuracies of all the five techniques used for prediction of attrition. Interestingly, all the five prediction techniques have shown reasonable accuracy levels with respect to those employees who remained with the company. On the other hand, discriminant analysis had given the highest accuracy level with respect to those who had left the company. Artificial neural networks are the lowest on predictive accuracy with respect to those who had left the company. From the company's perspective, it is important to predict those who are likely to leave the company more accurately so that pro-active strategies could be initiated to minimise the turnover levels. The company would be in a position to engage those who are predicted as likely to be leaving the company to identify the possible reasons even before the employees have made the final decision. On the whole the classification trees and CART appear to give the best results in terms of prediction accuracy. Both these techniques are able to predict the turnover with an accuracy level of above 80 per cent. At the same time the accuracy levels of these techniques with respect to those who remained with the company are above 90 per cent.

DISCUSSION

The findings of the five prediction models indicate that absenteeism and lateness, job content, demographics and experience in the current team (as one indicator of tenure) are strong predictors of turnover. There are unique context specific interpretations of these findings. Both the job content indicators namely type of position and the type of domain expertise emerged in the models as significant variables. There are two facets to this variable. The software services industry in India caters to global clients in a variety of sectors like financial and banking services, manufacturing, retail, and engineering. They also develop and maintain software developed on proprietary and standardised systems. In recent years, a number of product companies develop significant part of their products from their software development centres in India. Therefore, most large Indian and multinational organisations work in a variety of technologies and domains. The challenge is that most young engineers believe it is necessary to acquire experience on a variety of platforms and domains. Specialisation is currently not valued. Therefore, job content has meant working on a variety of platforms or domains, and it is possible that if another organisation offers a better domain or technology platform it will attract staff from organisations that offer less favourable platforms.

Secondly, it is well recognised that in rapid growth environments where demand for employees far outstrips the supply of labour, employees' expectations from jobs are very high. There could be a case of unmet expectations of the job, lack of challenging work or mismatch between skills and job content. Banerjee (2008) found that new recruits might have unrealistic expectations based on an image of an organisation that can turn into disillusionment if everyday reality does not match the expectations, leading to frustration and low productivity and eventually turnover. In the context of Indian software organisations, where a large number of projects are executed for overseas clients, there could be a mismatch between the skills required for the project and the expectations of the employee about where he/she should be placed. When a new project comes in, the employees from other completed projects or the bench are assigned. These employees may not necessarily have an interest or a liking for the technology. In particular, if the job requires working on a legacy technology, the employees may favour exit. Lastly, an attractive explanation of the above findings is that after several months/years of experience a person seeks for new job challenges, and, therefore, leaves the organisation if the company does not offer the employee a challenging job within the company. IT professionals tend to rate career development and a challenging job as greater than monetary compensation in determining their job satisfaction (Snyder, Rupp & Thornton 2006). Given the choice between money and a challenging job, many employees still prefer the latter as it allows them an opportunity to broad base their domain expertise and also provides an opportunity to work with cutting edge technology (Aravamudhan 2008). Therefore, a perceived lack of career development opportunities or unchallenging jobs have potential to cause turnover of software engineers, even though the company might pay high compensation.

The demographical variables, age and marital status are strong predictors of employee turnover. The results show that younger employees are more likely to turnover than older employees in early stages. These findings are in line with a study on voluntary turnover rates conducted by Hill and Associates which found that young undergraduates, graduates and post graduates in the outsourcing business had changed their jobs at least once in the past three years (Banerjee 2008). The role of age as a variable in the Indian context is particularly significant. The IT industry has been hiring a large number of young professionals from the campus. Most of these graduates are first time workforce entrants who appear to have unrealistic expectations from the job and the organisation. This unrealistic expectation coupled with scarcity of employable skills, and soaring salaries make them particularly vulnerable for turnover. It is also likely that many employees engage in a process of career exploration in their first few jobs. Existing research on careers also suggests that older employees given their career and life stage, may not be able to move as easily. Marital status as an indicator requires some additional explanation.

There are two factors which could explain this finding. In the Indian context, where duty to the family is of great importance, many professionals settle in cities which are closer to their home towns where their parents stay and often this occurs post marriage. With an increasing number of dual career couples and with child care support facilities being poor, families often become the source of child support. In the case of women, many of them change and move to cities where their spouses work. All these factors could contribute to marital status emerging as a significant predictor.

A strong predictor of employee exit is experience in the current team. The literature on Person-Group fit (P-G fit), mentions the importance of group member acceptance and engagement as a variable for a person's stay in an organisation (Kristof-Brown, Zimmerman & Johnson 2005). Employees can feel isolated and unhappy if they are not part of a cohesive team. As individuals' attitudes and behaviours will be influenced by the degree of congruence or 'fit' between individuals and organisations (Argyris 1957, Pervin 1989), it is likely that employees who feel that they do not fit in with their team members are more likely to show withdrawal behaviours and actually leave the organisation. This is supported by the findings that indicate casual leave and late arrival to be strong antecedents of employee turnover.

The frequency of job change as a predictor of turnover should be understood in the current labour market conditions where the demand for engineers far outstrips the supply. In such a context, it is possible for a young engineer to find job very easily. When 'reality shock' sets in, many employees tend to move to another organisation with a hope that the next organisation is going to be better. Such job shifts is common in early years, since many of them also do not have a clear picture of their career goals, aspirations and objectives.

CONCLUSION

This study investigated the relation between demographics, tenure, job content, frequency of job changes, withdrawal behaviours and employee turnover, using data mining techniques. 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 four variables namely demographics, tenure, job content and withdrawal behaviours 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.

AUTHORS

Vishnuprasad Nagadevara is a Professor in the area of Quantitative methods and Information Systems at the Indian Institute of Management Bangalore. His areas of research are Data Mining, applications of operations research techniques and Business Analytics.
Email: nagadev@iimb.ernet.in
Vasanthi Srinivasan is an Associate Professor in the area of Organizational Behaviour and Human Resource management. Her areas of interest are careers, Women and Technology, International Human Resource management and ethics and leadership. She has been a part of the team, which has been engaged in a European Union funded research "EMERGENCE" on understanding the global work relocations. She has extensive experience in the Information Technology sector in India and is engaged in capability building programs for HR professionals in South Asia.
Email: vasanthi@iimb.ernet.in

Reimara Valk is a Senior Research Analyst for Summit HR based at the Indian Institute of Management Bangalore (IIMB), India. Her areas of research are work-family balance of women and men IT professionals in India and staffing strategies of MNC companies.
Email: reimara.valk@iimb.ernet.in

ACKNOWLEDGEMENT

An earlier version of this paper was presented at the International Kyoto Conference at Kyoto, October 2007. The authors thank the participants of the conferences for their useful comments on the paper.

REFERENCES

Adler, S., & Golan, J. (1981). Lateness as withdrawal behaviour. *Journal of Applied Psychology*, 66, 544-554.

Ahuja, M. K., Chudoba, K. M., & Kacmar, C. J., McKnight, D. H., & George, J.F. (2007). IT road warrior: Balancing work–family conflict, job autonomy and work overload to mitigate turnover intentions. *MIS Quarterly*, 31(1), 1-17.

Aquino, K., Griffeth, R.W., Allen, D.G., & Hom, P.W. (1997). Integrating justice constructs into the turnover process: A test of a referent cognitions model. *Academy of Management Journal*, 40(5), 1208–1227.

Aravamudhan, N.R. (2008). Employee attrition: A costly dilemma for the organisation. *HRM Review*, (March), 59-63.

Argyris, C. (1957). The individual and organization: Some problems of mutual adjustment. *Administrative Science quarterly*, 2(1), 24.

Arora, A., & Athreye, S. (2002). The software industry and India's economic development. *Information Economics and Policy*, 14(2), 253-273.

Banerjee, I. (2008). Attrition: From corporate nightmare to competitive advantage. *HRM Review*, 68-72.

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

Bartlett, C.A., & Ghoshal, S. (1997). The myth of the generic managers: new personal competencies for new management roles. *California Management Review*, 40(1), 92-116.

Blau, G. (1994). Developing and testing taxonomy of lateness behaviour. *Journal of Applied Psychology*, 79(6), 959-970.

Booth, S., & Hamer, K. (2007). Labour turnover in the retail industry: Predicting the role of individual, organisational and environmental factors labour turnover in the retail industry. *International Journal of Retail & Distribution Management*, 35(4), 289-307.

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

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

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

Clegg, C. (1983). Psychology of employee lateness, absence and turnover. *Journal of Applied Psychology*, 68(1), 88-101.

Dess, G.D., & Shaw, J.D. (2001). Voluntary turnover, social capital, and organizational Performance. *Academy of Management Review,* 26(3), 446-456.

Hall, D. T. (2002). *Careers in and out of organizations.* Sage: USA

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.

Johns, G. (1995). Absenteeism. In N. Nicholson (Ed.), *The blackwell encyclopaedic dictionary of organizational behaviour* (1-3). Oxford, UK: Blackwell.

Kirschenbaum, A., & Mano-Negrin, R. (1999). Underlying labour market dimensions of "opportunities": The case of employee turnover. *Human Relations,* 52(10), 1233-1255.

Kozlowski, S.W.J., & Salas, E. (1997). An organizational systems approach for the implementation and transfer of training. In J.K. Ford, S.W.J. Kozlowski, K. Kraiger, E. Salas, & M. Teachcut (Eds.), *Improving training effectiveness in work organizations* (247-287). Mahwah, N.J.: Erlbaum.

Krausz M., Koslowsky., M. & Eiser, A. (1998). Distal and proximal influences on turnover intentions and satisfaction: Support for a withdrawal progression theory. *Journal of Vocational Behaviour*, 52(1), 59-71.

Kristof-Brown, A.L., Zimmerman, R.D., & Johnson, E.C. (2005). Consequences of individuals' fit at work: A meta-analysis of person-job, person-origination, person-group and person-supervisor fit. *Personnel Psychology,* 58(2), 281-342.

Lee T. W (1997). Employee turnover in L. H. Peters, C. R., Greer & S. A. Youngblood (Eds.), *The blackwell encyclopedic dictionary of human resource management* (97-100) Oxford Blackwell

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.

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

Lucas, G. H. Jr., Parasuraman, A., Davis, R. A., & Enis, B. M. (1987). An empirical study of salesforce turnover. *Journal of Marketing,* 51(3), 34-59.

Mano-Negrin, R. (2001). An occupational preference model of labour turnover. The case of Israel's medical sector employees. *Journal of Management in Medicine*, 15(2), 106-124.

Martocchio, J. J., & Harrison, D. A. (1993). To be there or not to be there? Questions, theories and methods in absenteeism research. *Research in Personnel and Human Resources Management,* 11(2), 259-328.

McLaughlin, K.J. (1991) A theory of quits and layoffs with efficient turnover. *Journnal of Political Economy,* 99(1), 1-29.

Mitra, A., Jenkins, G .D. Jr., & Gupta, N. (1992). A Meta-analytic review of the relationship between absence and turnover. *Journal of Applied Psychology*, 77(60), 879-889.

Mobley, W.H. (1977). Intermediate linkages in relationship between job satisfaction and employee turnover. *Journal of Applied Psychology*, 62(2), 237–240.

Moitra, D. (2001). India's software industry. *IEEE Software,* Jan-Feb, 77-80.

Moore, J. E., & Burke, L. (2002). How to turn around "turnover culture" in IT. *Communications of the ACM,* 45(1), 73-78.

Morrell, K. Loan-Clarke J., & Wilkinson, A. (2004). The role of shocks in employee turnover. *British Journal of Managemen*t, 15(3), 335-349.

Nagadevara, V. (2004). *Application of neural prediction models in healthcare*. Paper presented at the Second International Conference on e-Governance, Colombo, Sri Lanka.

Nagadevara, V. (2008). Improving the effectiveness of the hotel loyalty programs through data mining. In V. Jauhari (Ed.), Global causes on hospitality industry, New York: The Haworth Press.

NASSCOM. (2006). HR initiatives, HR best practices in global sourcing. [Online] NASSCOM Newsline Issue No. 59, September 2006. Available: http://www.nasscom.in/Nasscom/templates/NormalPage.aspx?id=50205 [2007, March 1].

Pervin, L.A. (1989). Persons, situations, interactions: The history of a controversy and a discussion of theoretical models. *Academy of Management Review*, 14(3), 350 –360.

Pfeffer, (1994) *Competitive advantage through people: Unleashing the power of the work force.* Boston: Harvard Business School Press.

Porter, L.W., Crampion, W.J., & Smith, F.J. (1976). Organizational commitment and managerial turnover: A longitudinal study. *Organizational Behavior and Human Performance*, 15(1), 87–98.

Rajawat, K.Y. (2007). Taking on the big mouse. *Businessworld*, 36-42.

Ramani, V.V., & Raghunandan, U.N. (2008). Managing attrition level in organizations. *HRM Review*, 33-38.

Rathi, N. S. (2003). Human resources challenges in Indian software industry: An empirical study of employee turnover. Thesis PhD degree. Shailesh J. Mehta School of Management, Indian Institute of Technology, Bombay.

Rosse, J. G. (1988). Relations among lateness, absence, and turnover: Is there a progression of withdrawal? *Human Relations*, 41(4), 517-531.

Sharma, S. (2007). "High attrition rate: A big challenge", [Online] http://www.bpoindia.org/research/attrition-rate-big-challenge.shtml [2007, 31 March].

Shaw, J. D., Delery. J. E., Jenkins, G. D., & Gupta, N. (1998). An organization-level analysis of voluntary and involuntary turnover. *Academy of Management Review*, 41(5), 511–525.

Snyder, L.A., Rupp, D.E., & Thornton, G.C. (2006). Personnel selection of information technology workers: The people, the jobs and issues for human resource management. *Personnel and Human Resource Management*, 25(3), 305-376.

Somers, M. J. (1999). Application of two neural network paradigms to the study of voluntary employee turnover. *Journal of Applied psychology*, 84(1), 177-185.

Steel, R.P., & Rentsch, J.R. (1995). Influence of cumulation strategies on the long range predictions of absenteeism. *Academy of Management Journal*, 20(5), 1616 -1634.

Tett, R.P., & Meyer, J.P. (1993.) Job-satisfaction, organizational commitment, turnover intention, and turnover: Path analyses based on meta-analytic findings. *Personnel Psychology*, 46, 259–293.

Wanous, J.P. (1992). *Organizational entry*. Reading, MA: Addison-Wesley.

Wolpin, J., Burke, R. J., Krausz, M., & Freibach, N. (1988). Lateness and absenteeism: An examination of the progression hypothesis. *Canadian Journal of Administrative Sciences*, 5(1), 49-54.

Table 1
Relative Importance of the Inputs Used in the Construction of ANNs

Variable	Relative Importance
Privilege leave used (Binned)	0.148577
Current position in the company (Binned)	0.146736
Current technical expertise	0.146736
Late arrival at the office (Binned)	0.144292
Experience in the company (Binned)	0.082996
Total experience (Binned)	0.078281
Domain experience	0.071921
Casual leave used (Binned)	0.069042
Total number of job changes	0.034097
Marital status	0.025611
Gender	0.014790
Age (Binned)	0.010350
Experience in the current position (Binned)	0.004343
Experience in the current team (Binned)	0.003620

Table 2
Variables in the Equation

Variable	B	Exp(B)
Age (in years)	.762	2.142
Gender	-.770	.463
Marital status	1.699	5.470
Total experience in years	-.984	.374
Experience at the company	.263	1.300
Experience in the current Team	-.461	.631
Experience in the current position	-.334	.716
Total number of job changes	-.459	.632
Casual leave used	.410	1.506
Privilege leave used	-.119	.888
Late arrival at the office	-.262	.769
Current Position (Binned)		
Current position binned(1)	-.828	.437
Current position binned(2)	.623	1.865
Current position binned(3)	-.409	.664
Current position binned(4)	-2.318	.098
Current position binned(5)	3.612	37.035
Constant	-14.489	.000

Table 3
Standardised Canonical Discriminant Function Coefficients

Variable	Coefficient
Age (in years)	-1.437
Total experience in years	1.815
Experience at the company	-.832
Experience in the current team	.629
Experience in the current position	.615
Current technical expertise	.002
Total number of job changes	.197
Casual leave used	-.306
Privilege leave used	.204
Late arrival at the office	-.073

Table 4
Prediction Accuracies of Different Techniques Used %

Status	Prediction		
	Left the Company	Not Left the Company	Total
	C 5.0		
Left the company	82.93	17.07	100.00
Not left the company	9.43	90.57	100.00
	CART		
Left the company	80.49	19.51	100.00
Not left the company	6.60	93.40	100.00
	Logistic Regression		
Left the company	75.00	25.00	100.00
Not left the company	18.63	81.37	100.00
	The cut value is 0.700		
	Artificial Neural Networks		
Left the company	58.54	41.46	100.00
Not left the company	9.43	90.57	100.00
	Discriminant Analysis		
Left the company	86.84	13.16	100.00
Not left the company	19.79	80.21	100.00

Figure 1
A Model Framework on the Relationship between Attrition and Behavioural Variables

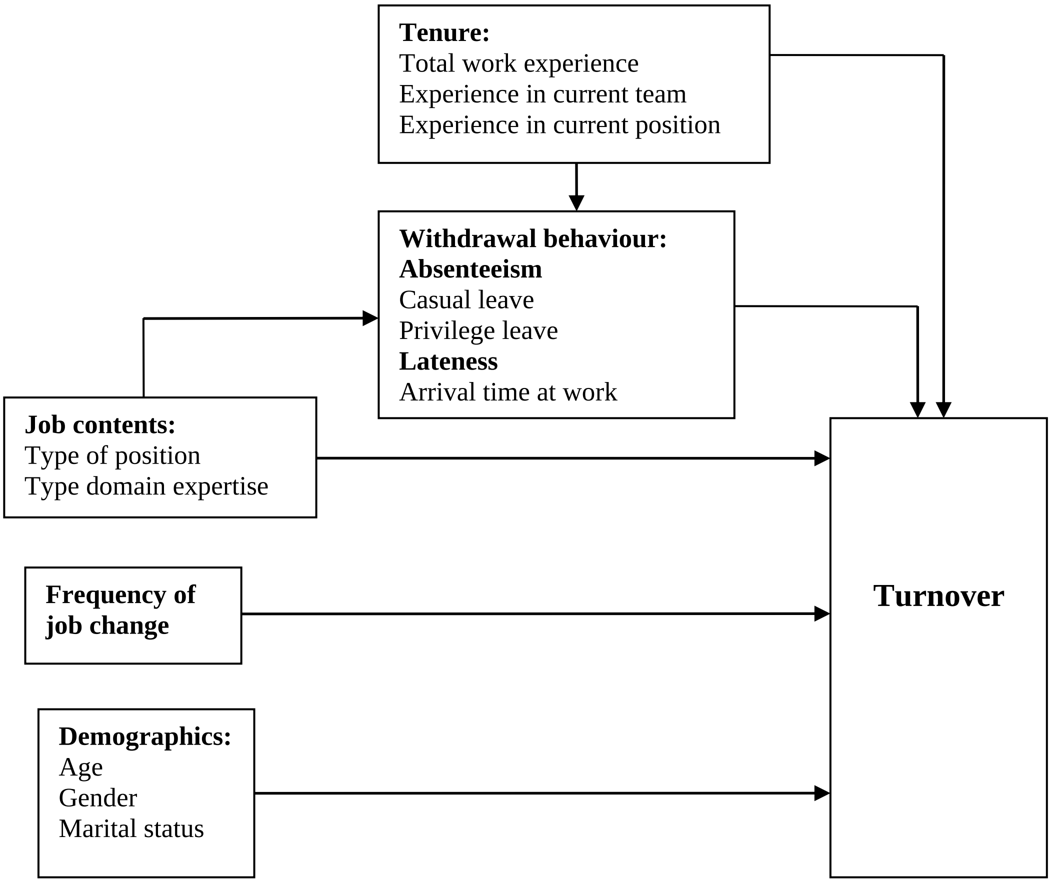


Figure 2
Classification and Regression Tree Diagram

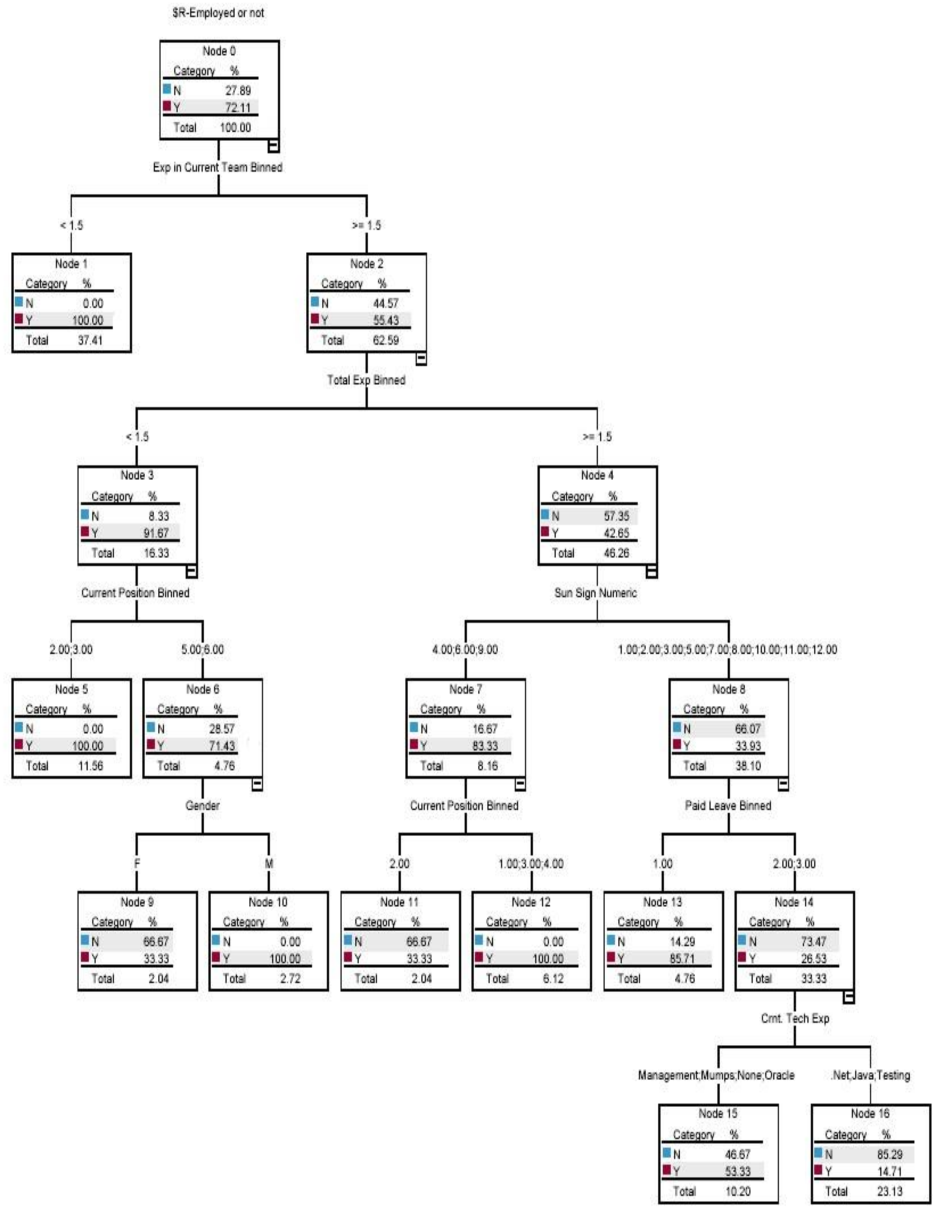
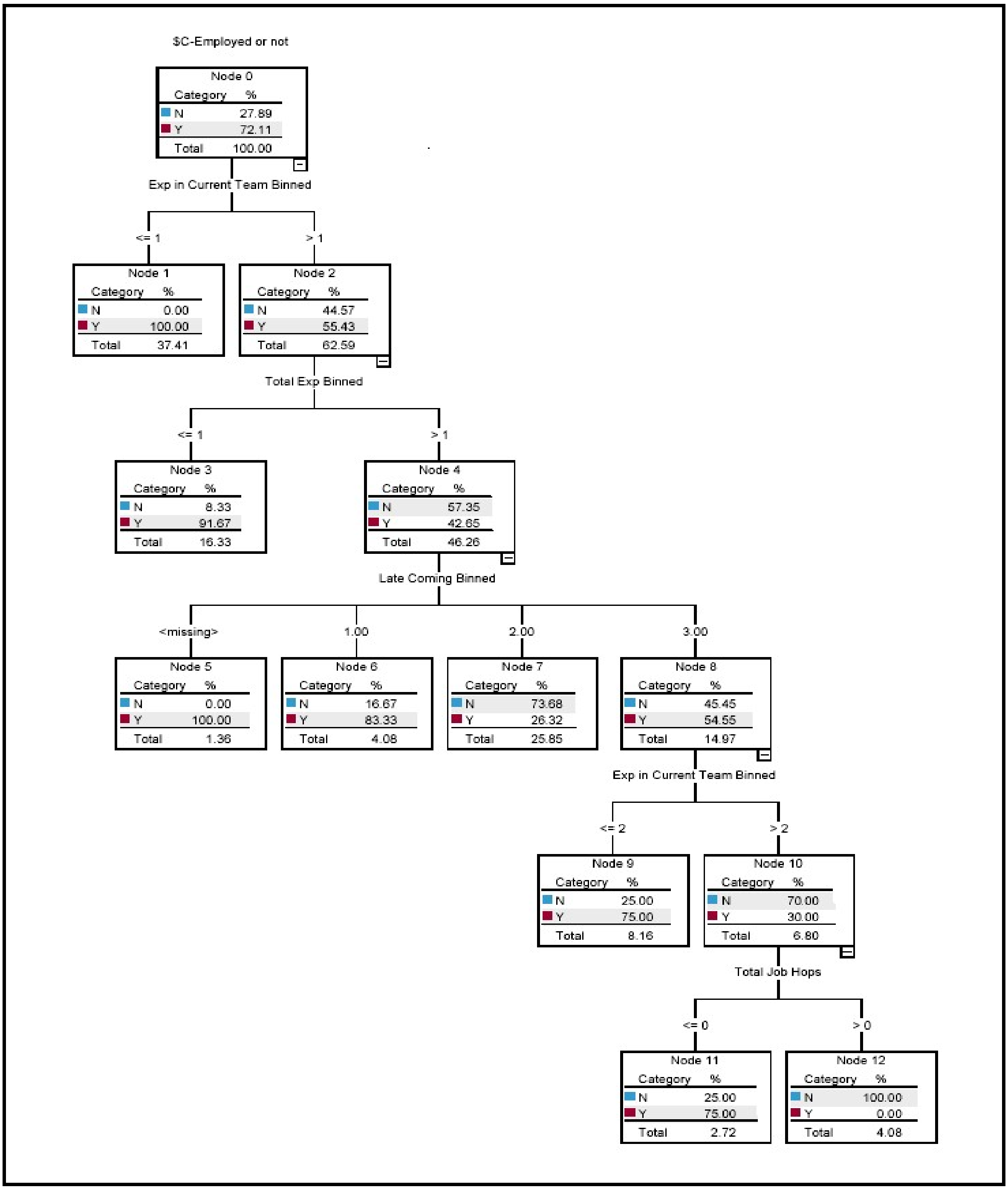


Figure 3
Classification Tree Diagram (C5.0)


$C-Employed or not
Node 0
Category %
N 27.89
Y 72.11
Total 100.00
Exp in Current Team Binned
<= 1
> 1
Node 1
Category %
N 0.00
Y 100.00
Total 37.41
Node 2
Category %
N 44.57
Y 55.43
Total 62.59
Total Exp Binned
<= 1
> 1
Node 3
Category %
N 8.33
Y 91.67
Total 16.33
Node 4
Category %
N 57.35
Y 42.65
Total 46.26
Late Coming Binned
<missing>
1.00
2.00
3.00
Node 5
Category %
N 0.00
Y 100.00
Total 1.36
Node 6
Category %
N 16.67
Y 83.33
Total 4.08
Node 7
Category %
N 73.68
Y 26.32
Total 25.85
Node 8
Category %
N 45.45
Y 54.55
Total 14.97
Exp in Current Team Binned
<= 2
> 2
Node 9
Category %
N 25.00
Y 75.00
Total 8.16
Node 10
Category %
N 70.00
Y 30.00
Total 6.80
Total Job Hops
<= 0
> 0
Node 11
Category %
N 25.00
Y 75.00
Total 2.72
Node 12
Category %
N 100.00
Y 0.00
Total 4.08