PREDICTING EMPLOYEE ATTRITION WITH TEN OR MORE YEARS OF EMPLOYMENT THROUGH CLASSIFICATION ALGORITHM

Abbas Heiat, Montana State University-Billings, 1500 University Drive, 59101, 406-657-1627,

aheiat@msubillings.edu

ABSTRACT

The purpose of this study is to investigate individual employee characteristics and organizational variables that may lead to attrition of Employee with 10 or more years of employment. C5 classification methods used to develop models for predicting employee attrition. The training model has a high accuracy at 96.25 % while testing model has only 59% of accuracy. However, some of the findings of this study in terms of important predicting variables are different from previous studies"

INTRODUCTION

Many researchers have indicated that the most valuable asset and important resource in organizations are their employees and employee attrition is considered to be a serious issue for organizations [1]. The cost of replacing employees is very high. Organizations need to search, hire and train new employees. Loss of experienced workers especially high performers is difficult to manage and is negatively related to the success and performance of organizations [2, 3, 4, 5, 6, 7, and 8]. The purpose of this study is twofold. First, to investigate individual employee characteristics and organizational variables that may lead to employee turnover. Identifying the most relevant factors influencing employee attrition is essential for

implementing business strategies by selecting and adjusting proper improvement activities for retaining and hiring new employees. Second, to develop a model for predicting probable employee attrition by using data mining algorithms. In my previous study two classification methods used to develop models for predicting employee attrition. Artificial Neural Network (ANN) model predicted the employee attrition more accurately (85.33%) than Decision Tree (C&R Tree) model (80.89%). Both models, however, determined years at the company and working overtime as the most important variables influencing employee attrition. This is in contrast to findings in previous findings that job satisfaction, security, training rewards and employee participation in organizational decision-making are most influential variables. In this study only the data for the employees with ten or more years of employment used to determine contributing factors to employee attrition [17].

DATA PROCESSING

The data used in this research provided by IBM Watson Analytics Community-Human Resource Employee Attrition. Data Included 36 variables including the dependent variable attrition.

To analyze the data categorical variables needed to be preprocessed for data mining. Certain variables had to be taken into account and others excluded. The excluded variables did not have any likely impact on the employee attrition. Only data for employees with ten or more years of employment is included in this study. The data was prepared and run through exploratory analysis which in Modeler is called Feature Selection in order to find the most influential variables. The data was doctored to help fill the gaps with the missing data. The data was then broken into training and test/validation sets to develop the model(s) and validate the results of analysis. Since the percentage of records which indicated employee attrition is low compared to non-attrition records, a balance node is used to make proportions of attrition and non-attrition almost the same. The target or dependent variable is employ attrition.

METHODOLOGY

Data Mining may be defined as the process of finding potentially useful patterns of information and relationships in data. As the quantity of clinical data has accumulated, domain experts using manual analysis have not kept pace and have lost the ability to become familiar with the data in each case as the number of cases increases. Improved data and information handling capabilities have contributed to the rapid development of new opportunities for knowledge discovery. Interdisciplinary research on knowledge discovery in databases has emerged in this decade. Data mining, as automated pattern recognition, is a set of methods applied to knowledge discovery that attempts to uncover patterns that are difficult to detect with traditional statistical methods. Patterns are evaluated for how well they hold on unseen cases. Databases, data warehouses, and data repositories are becoming ubiquitous, but the knowledge about the relationship among variables are still lacking. The most efficient algorithms with highest accuracy rates were C5 based on current data set used for analysis.

RESULTS OF ANALYSIS

The following figure shows the importance of input variables for predicting attrition. This is in contrast to findings in literature which indicate that job satisfaction, security, training rewards and employee participation in organizational decision-making are most influential variables.

Confusion matrix in Figure 5 shows that the C5 model using the training dataset is 96.25 % accurate in classifying attrition while using the test (validation) dataset predicts employee attrition correctly 58.94 % of the time. The gain charts in Figure 6 demonstrate the improvement gained by using C5 model as compared with a non-model approach like using average attrition.

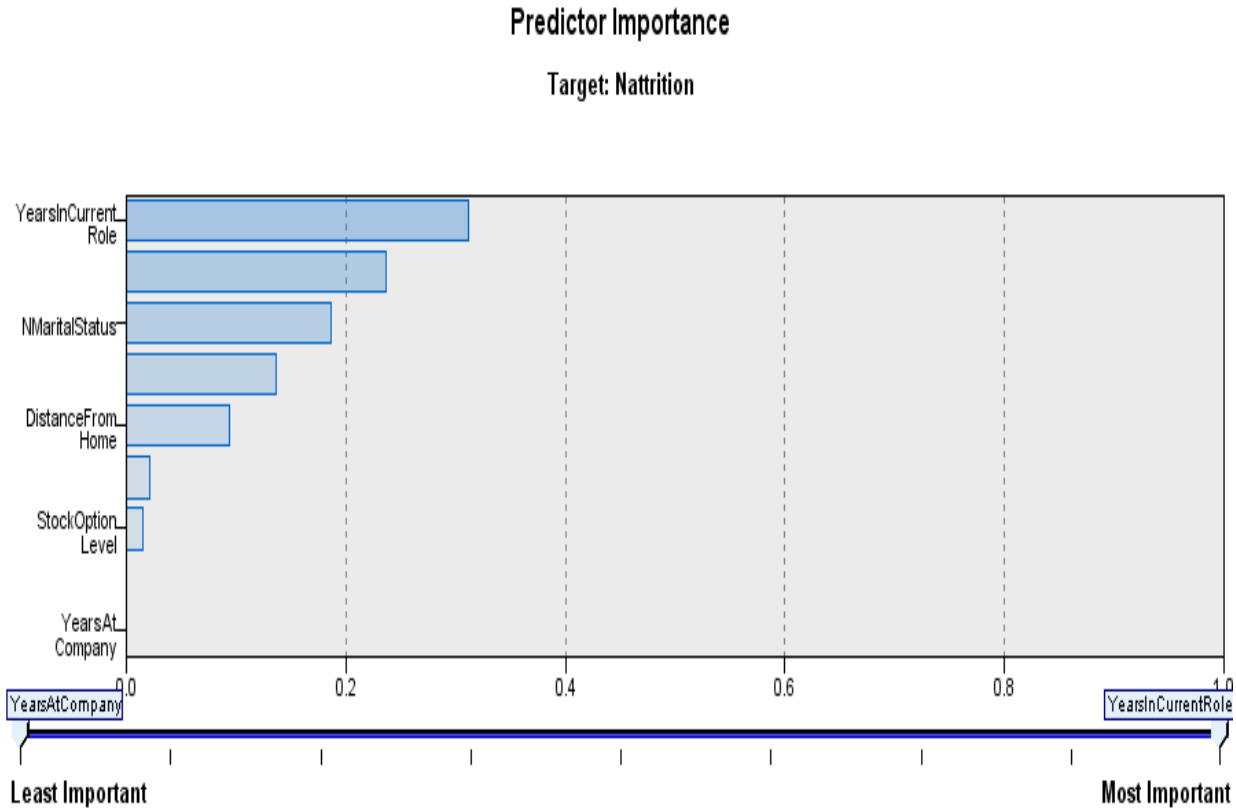

Figure 1. Importance of Variable According to Decision Tree

Comparing $C-Nattrition with Nattrition

'Partition'	1_Training		2_Testing	
Correct	487	96.25%	122	58.94%
Wrong	19	3.75%	85	41.06%
Total	506		207	

Coincidence Matrix for $C-Nattrition (rows show actuals)

'Partition' = 1_Training	0.000000	1.000000
0.000000	244	19
1.000000	0	243

'Partition' = 2_Testing	0.000000	1.000000
0.000000	68	13
1.000000	72	54

Figure 2. Decision Tree Confusion Matrix

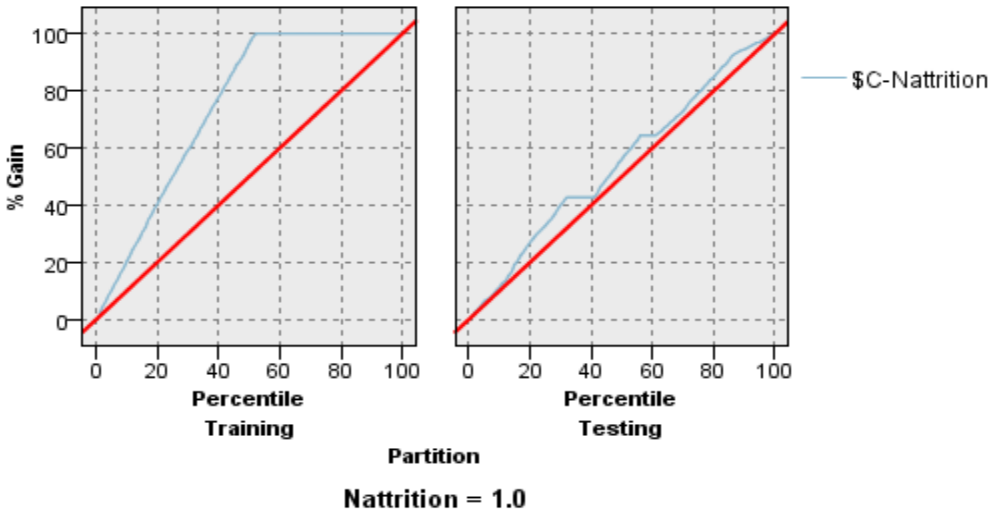

Figure 3. C5 Gain Chart

CONCLUSION

C5 classification methods used to develop a model for predicting employee attrition. C5 model determined that years at the current role, marital status, distance from home and stock option as the most important variables influencing employee attrition. This is in contrast to some of the previous findings that job satisfaction, security, training rewards and employee participation in organizational decision-making are most influential variables. One surprising result is that the numbers of the years at a company does not seem to be important in employees' decisions to stay or leave the company.

The contrasting results might be due to different type and environment of organizations about them the data were collected. In that case separate models based on different types and environments should be developed. Further studies are needed to investigate, confirm or reject the validity of the last statement.

REFERENCES ARE AVAIALBLE UPON REQUEST