# PREDICTING EMPLOYEE ATTRITION THROUGH DATA MINING

*Abbas Heiat, College of Business, Montana State University, Billings, MT 59102,*
*aheiat@msubillings.edu*

## ABSTRACT

The purpose of this study is to investigate individual employee characteristics and organizational variables that may lead to employee attrition. 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%). However, 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.

## 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. 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. 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 and skills required to capitalize on these collections of data are not yet widespread. In this research As a First step I used Auto-Classification tool in SPSS Modeler which applies 11 different algorithms. The most efficient algorithms with highest accuracy rates were C&R Tree and Neural Net based on current data set used for analysis. The following is a brief description of the algorithms suggested by Auto-Classification as the most accurate models for our dataset.

**Decision Trees**- Decision trees and rule induction are two most commonly used approaches to discovering logical patterns within data sets. Decision trees may be viewed as a simplistic approach to rule discovery because of the process used to discover patterns within data sets. Decision tree is built through a process known as binary recursive partitioning. This is an iterative process of splitting the data into partitions, and then splitting it up further on each of the branches. Initially, you start with a training set in which the classification label (say, "attrition" or "no-attrition ") is known (pre-classified) for each record. All of the records in the training set are together in one big box. The algorithm then systematically tries breaking up the records into two parts, examining one variable at a time and splitting the records on the basis of a dividing line in that variable (say, age> 30 or age<=30). The object is to attain as homogeneous set of labels (say, "attrition" or "no-attrition") as possible in each partition. This splitting or partitioning is then applied to each of the new partitions. The process continues until no more useful splits can be found. The heart of the algorithm is the rule that determines the initial split rule [14].

The process starts with a training set consisting of pre-classified records. Pre-classified means that the target field, or dependent variable, has a known class or label: "productive" or "non-productive". The goal is to build a tree that distinguishes among the classes. For simplicity, assume that there are only two target classes and that each split is binary partitioning. The splitting criterion easily generalizes to multiple classes, and any multi-way partitioning can be achieved through repeated binary splits. To choose the best splitter at a node, the algorithm considers each input field in turn. In essence, each field is sorted. Then, every possible split is tried and considered, and the best split is the one which produces the largest decrease in diversity of the classification label within each partition. This is repeated for all fields, and the winner is chosen as the best splitter for that node. The process is continued at the next node and, in this manner, a full tree is generated.

**Artificial Neural Networks (ANN)** - Artificial neural networks are defined as information processing systems inspired by the structure or architecture of the brain. They are constructed from interconnecting processing elements, which are analogous to neurons. The two main techniques employed by neural networks are known as supervised learning and unsupervised learning. In unsupervised learning, the neural network requires no initial information regarding the correct classification of the data it is presented with. The neural network employing unsupervised learning is able to analyze a multi-dimensional data set in order to discover the natural clusters and sub-clusters that exist within that data. Neural networks using this technique are able to identify their own classification schemes based upon the structure of the data

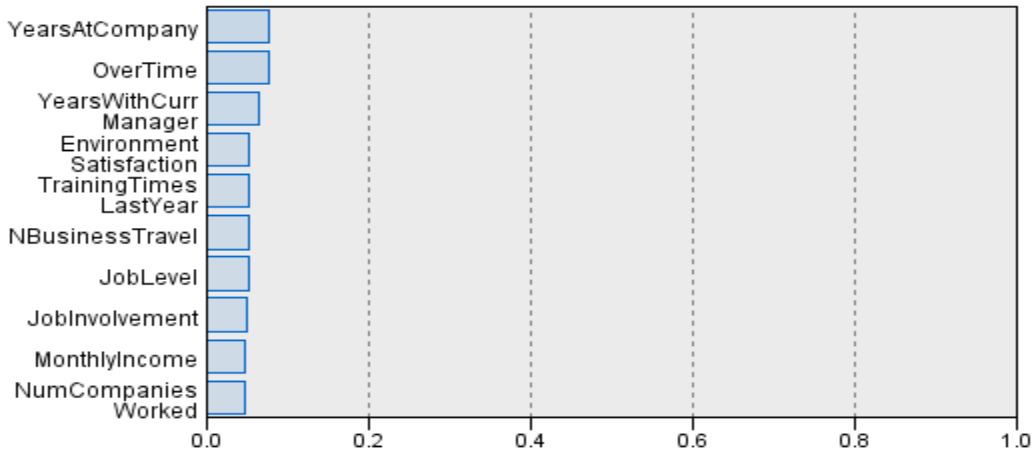
provided, thus reducing its dimensionality. Unsupervised pattern recognition is therefore sometimes called cluster analysis [15, 16]. Supervised learning is essentially a two stage process; firstly training the neural network to recognize different classes of data by exposing it to a series of examples, and secondly, testing how well it has learned from these examples by supplying it with a previously unseen set of data. A trained neural network can be thought of as an "expert" in the category of information it has been given to analyze. It provides projections given new situations of interest and answers "what if" questions. There are disadvantages in using ANN. No explanation of the results is given i.e. difficult for the user to interpret the results. They are slow to train due to their iterative nature. Empirical studies have shown that if the data provided does not contain useful information within the context of the focus of the investigation, then the use of neural networks cannot generate such information any more than traditional analysis techniques can. However, it may well be the case that the use of neural networks for data mining allows this conclusion to be reached more quickly than might ordinarily be the case.

## ANLYSIS RESULTS

As the following diagram demonstrates Auto-Classifier tool was used to arrive at the best model for employee attrition. The diagram starts with selecting the data set for the analysis. It follows with a Filter and Type node that selects the appropriate inputs and assigns the appropriate data type to the target and input variables. Since the dataset was quite large, the records with missing values were discarded. Next, the cleaned dataset was partitioned to training and testing sets (70%, 30%). Once the Auto-classifier determined the best model, that algorithm was applied to the dataset and analysis and evaluations nodes were added to analyze the results. Figure 3 SPSS Modeler created models for employee attrition dataset.

**ANN Results**- Based on the neural network algorithm the most important Variables are shown in Figure 4. According to neural network analysis Years at the company, Working Overtime, and Years with Current Manger are the most important factors influencing the decision of an employee to stay or to leave an organization. Other less important variables are listed in Figure 4. 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.

**Figure 2. Importance of Variables According ANN analysis**



Confusion matrix in Figure 5 shows that the ANN model using the test (validation) dataset predicts employee attrition correctly 85.33% of the time. The gain charts in Figure 6 demonstrate the improvement gained by using ANN model as compared with a non-model approach like using average attrition.

**Figure 3. ANN Confusion Matrix**

Comparing $N-Nattrition with Nattrition

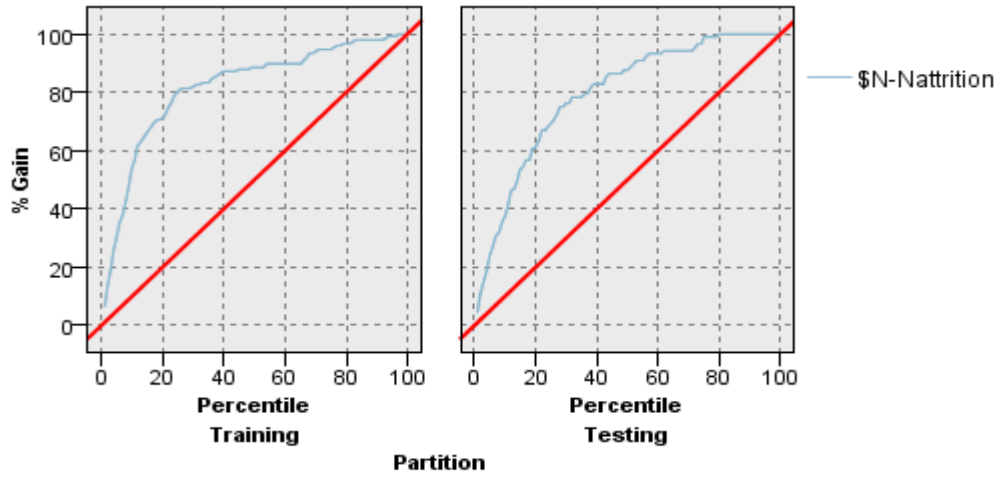
| 'Partition' | 1_Training | | 2_Testing | |
|---|---|---|---|---|
| Correct | 922 | 90.39% | 384 | 85.33% |
| Wrong | 98 | 9.61% | 66 | 14.67% |
| Total | 1,020 | | 450 | |

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

| 'Partition' = 1_Training | 0.000000 | 1.000000 |
|---|---|---|
| 0.000000 | 852 | 19 |
| 1.000000 | 79 | 70 |

| 'Partition' = 2_Testing | 0.000000 | 1.000000 |
|---|---|---|
| 0.000000 | 352 | 10 |
| 1.000000 | 56 | 32 |

**Figure 4. ANN Gain Chart**




**Decision Tree Results-** Based on C&R Tree algorithm the most important Variables are shown in Figure 7. According to neural network analysis Working Overtime, Years at the company and Job Role are the most important factors influencing the decision of an employee to stay or to leave the organization. Other less important variables are listed in Figure 7. Once more this conclusion 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.

**Figure 5. Importance of Variable According to Decision Tree**

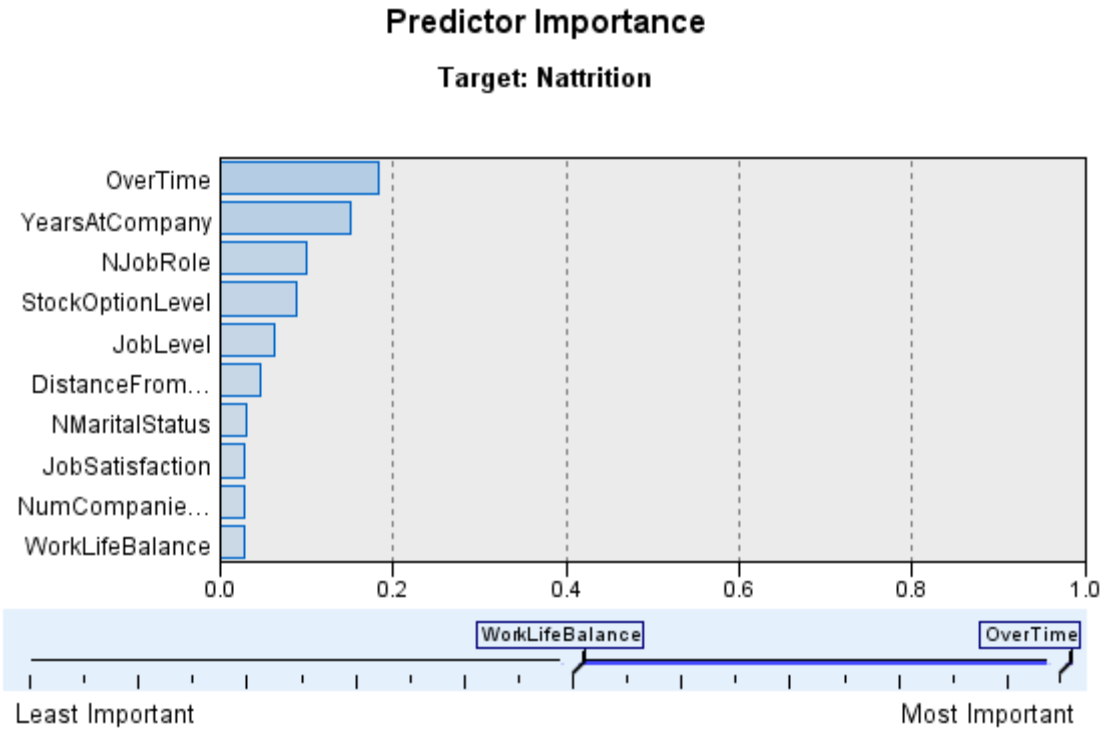



**Figure 6. Decision Tree Confusion Matrix**

Comparing $R-Nattrition with Nattrition

| 'Partition' | 1_Training | | 2_Testing | |
|---|---|---|---|---|
| Correct | 894 | 87.65% | 364 | 80.89% |
| Wrong | 126 | 12.35% | 86 | 19.11% |
| Total | 1,020 | | 450 | |

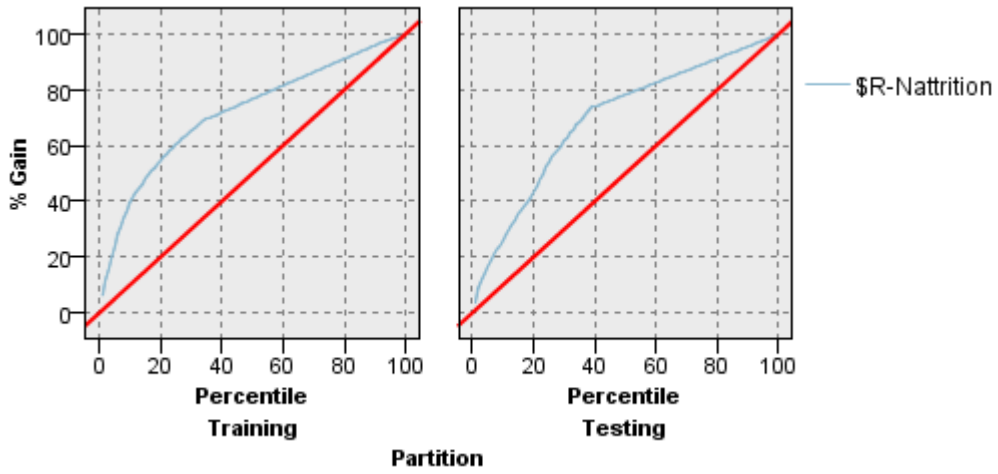
Coincidence Matrix for $R-Nattrition (rows show actuals)

| 'Partition' = 1_Training | 0.000000 | 1.000000 |
|---|---|---|
| 0.000000 | 850 | 21 |
| 1.000000 | 105 | 44 |

| 'Partition' = 2_Testing | 0.000000 | 1.000000 |
|---|---|---|
| 0.000000 | 343 | 19 |
| 1.000000 | 67 | 21 |

Confusion matrix in Figure 8 shows that the C&R Tree model using the test (validation) dataset predicts employee attrition correctly 80.89% of the time. The gain charts in Figure 9 demonstrate the improvement gained by using ANN model as compared with a non-model approach like using average attrition.

**Figure 7. C&R Tree Gain Chart**




## CONCLUSION

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.

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 available upon request.