PREDICTING AND COMPARING VARIABLESVCONTRIBUTING TO ATTRITION OF EMPLOYEES

Abbas Heiat, College of Business, Montana State University-Billings, University Drive, Billings, MT 59102, 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 less years of employment. C5 classification methods used to develop models for predicting employee attrition. The testing dataset has a high accuracy at 97.6 % for prediction employees' attrition and 93% accuracy for predicting non-attrition employees. Some of the findings of this study in terms of important predicting variables are different from previous studies.

INTRODUCTION

Employees are the most valuable asset and an important resource in many organizations. Therefore, employee attrition is a serious loss for organizations [1]. Hiring new employees and raining them would be difficult and cost a lot. Research indicates that the loss of experienced workers has a negative impact on the success and performance of organizations [2, 3, 4, 5, 6, 7, and 8]. The purpose of this study is twofold. First, to investigate, using data mining models, employee characteristics and organizational variables that may lead to recently hired 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 compare these relevant variables with important variables for attrition of employees with more than 10 years employment. In my previous study, C5 algorithm determined that years at the current role, marital status, distance from home, working overtime and stock option are the most important variables influencing employee attrition predicting employee attrition. The C5 using the test (validation) dataset predicts employee attrition correctly 58.94 % of the time. This was 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 less years of employment used to determine contributing factors to employee attrition [17].

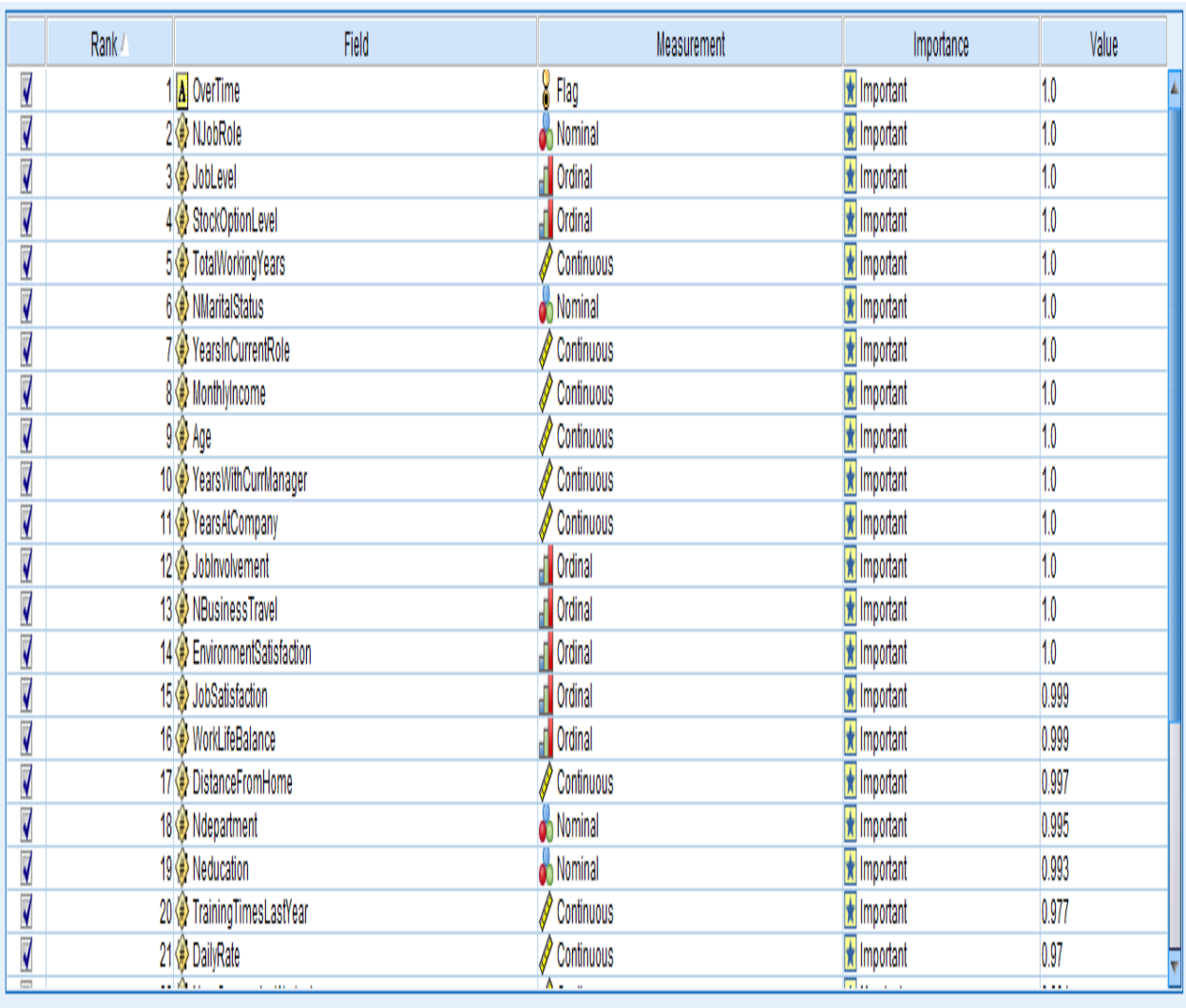
Data Processing

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 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 less years of employment is included in this study. The data was prepared and run through exploratory analysis, which in IBM SPSS Modeler is called Feature Selection, to find the most influential variables that have effect on dependent variable namely Attrition variable. 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. The following figure shows the variables included in the dataset.

Figure 1. Variables included in the analysis based on Feature Selection



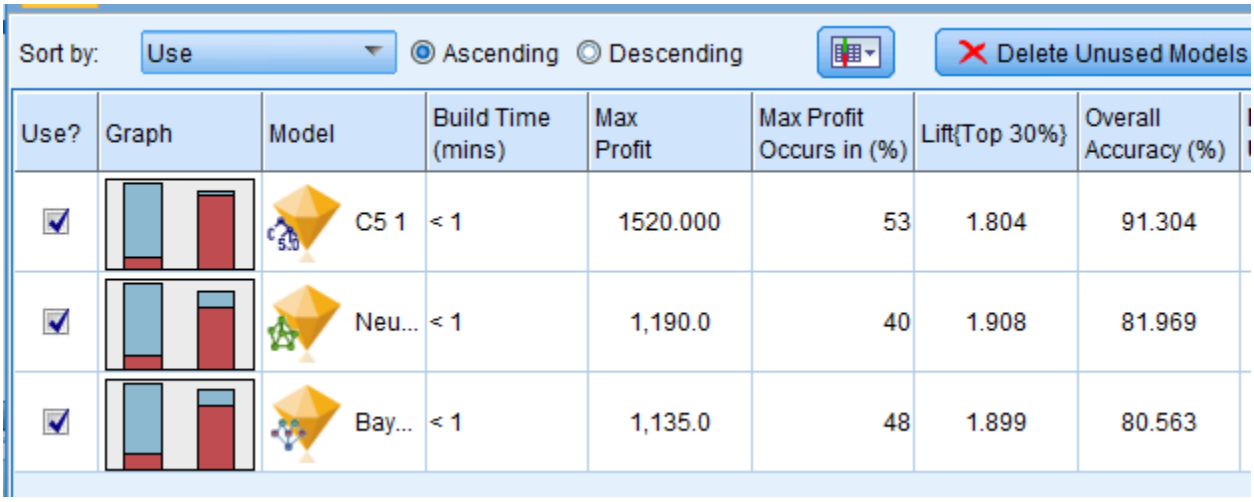
	Rank	Field	Measurement	Importance	Value
☑	1	OverTime	Flag	Important	1.0
☑	2	NJobRole	Nominal	Important	1.0
☑	3	JobLevel	Ordinal	Important	1.0
☑	4	StockOptionLevel	Ordinal	Important	1.0
☑	5	TotalWorkingYears	Continuous	Important	1.0
☑	6	NMaritalStatus	Nominal	Important	1.0
☑	7	YearsInCurrentRole	Continuous	Important	1.0
☑	8	MonthlyIncome	Continuous	Important	1.0
☑	9	Age	Continuous	Important	1.0
☑	10	YearsWithCurrManager	Continuous	Important	1.0
☑	11	YearsAtCompany	Continuous	Important	1.0
☑	12	JobInvolvement	Ordinal	Important	1.0
☑	13	NBusinessTravel	Ordinal	Important	1.0
☑	14	EnvironmentSatisfaction	Ordinal	Important	1.0
☑	15	JobSatisfaction	Ordinal	Important	0.999
☑	16	WorkLifeBalance	Ordinal	Important	0.999
☑	17	DistanceFromHome	Continuous	Important	0.997
☑	18	Ndepartment	Nominal	Important	0.995
☑	19	Neducation	Nominal	Important	0.993
☑	20	TrainingTimesLastYear	Continuous	Important	0.977
☑	21	DailyRate	Continuous	Important	0.97

Methodology

Data Mining may be defined as the process of finding potentially useful patterns of information and relationships in data. As the quantity of data has accumulated in databases, 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 the past 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. As a first step, I used Auto-Classification tool in IBM SPSS Modeler that applies 11 different algorithms shown in Figure 2 to determine the most appropriate models.

Figure 2. Auto-Classification's Algorithms



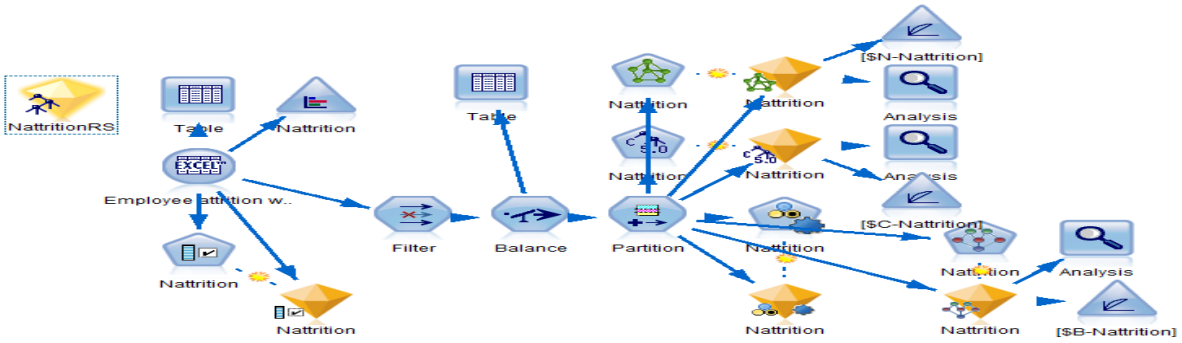
Use?	Graph	Model	Build Time (mins)	Max Profit	Max Profit Occurs in (%)	Lift{Top 30%}	Overall Accuracy (%)
☑		C5 1	< 1	1520.000	53	1.804	91.304
☑		Neu...	< 1	1,190.0	40	1.908	81.969
☑		Bay...	< 1	1,135.0	48	1.899	80.563

The most efficient algorithms with highest accuracy rates were C5 based on current data set used for analysis following Artificial Neural Network, and Bayes Net. In order to compare the results of this study with previous study, in which we used the most accurate algorithm namely C5, I am going to use C5 that is also the most accurate algorithm in the current study.

Analysis Results

The following IBM SPSS Modeler diagram starts with selecting the data set for the analysis. It follows with a Feature Selection Filter and Type node that selects the important input variables and assigns the appropriate data type to the target and input variables. Next, Distribution and the Balance Node is added to make proportions of attrition and non-attrition records almost equal. In the Next step, the dataset was partitioned to training and testing sets (70%, 30%). Then the Auto-classifier determined the best algorithms that algorithm was applied to the dataset. Analysis and Evaluations nodes were added to analyze the results. Figure 3 shows the IBM SPSS Modeler model for employee attrition dataset.

Figure 3. Models Created by SPSS Modeler



Decision Tree (C5)

I used the Decision Tree (C5) algorithm for analysis of the attrition dataset. Figure 4 shows the important variables that contribute to the employee attrition.

Figure 4. Important of Variables According to Decision Tree

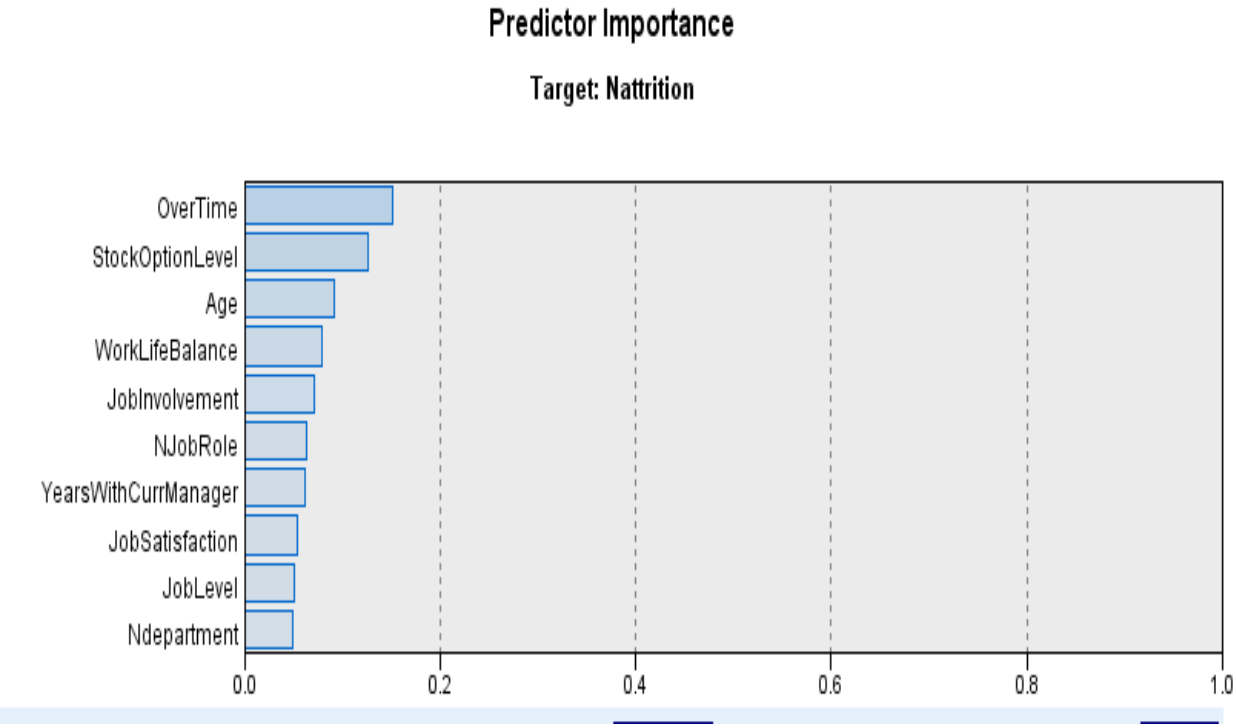


Figure 5. Decision Tree Confusion Matrix

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

'Partition' = 1_Training	0.000000	1.000000
0.000000	729	94
1.000000	12	848
'Partition' = 2_Testing	**0.000000**	**1.000000**
0.000000	381	29
1.000000	9	368

Confusion matrix in Figure 5 shows that the C5 model using the testing dataset is 97.6 % accurate in classifying attrition (1) while using the test dataset predicts non-attritional employees correctly 93% 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 6. C5 Gain Chart

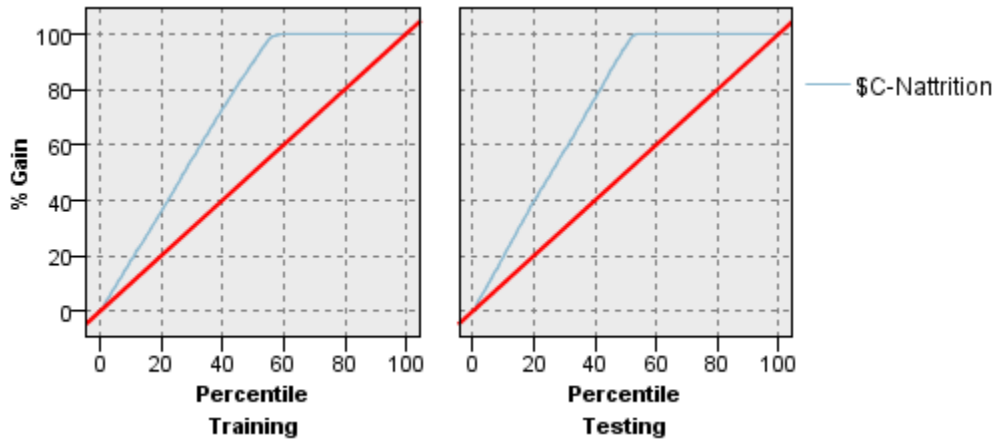


Figure 7. C5 Classification and Rule set showing the first three rules, Non-attrition

Rules for 0 - contains 77 rule(s)

Rule 1 for **0.0** (4; 0.75)
- if TotalWorkingYears <= 2
- and OverTime = No
- and Age <= 32
- and MonthlyIncome > 2,684
- and NJobRole in [2.000]
- then **0.000**

Rule 2 for **0.0** (14; 1.0)
- if TotalWorkingYears <= 2
- and OverTime = No
- and Age <= 32
- and DistanceFromHome <= 16
- and WorkLifeBalance in [1.000 2.000 3.000]
- and NJobRole in [3.000]
- then **0.000**

Rule 3 for **0.0** (3; 1.0)
- if TotalWorkingYears <= 2
- and OverTime = No
- and Age <= 32
- and DailyRate <= 1,041
- and EnvironmentSatisfaction in [4.000]
- and NBusinessTravel in [0.000 1.000]
- and NJobRole in [7.000]
- then **0.000**

Figure 8. C5 Classification and Rule set showing the first four rules for Attrition

Rules for 1 - contains 61 rule(s)

Rule 1 for **1.0** (95; 0.926)
- if TotalWorkingYears <= 2
- and OverTime = Yes
- then **1.000**

Rule 2 for **1.0** (38; 0.947)
- if TotalWorkingYears <= 2
- and OverTime = No
- and Age <= 32
- and MonthlyIncome <= 2,684
- and NJobRole in [2.000]
- then **1.000**

Rule 3 for **1.0** (5; 1.0)
- if TotalWorkingYears <= 2
- and OverTime = No
- and Age <= 32
- and DistanceFromHome > 16
- and WorkLifeBalance in [1.000 2.000 3.000]
- and NJobRole in [3.000]
- then **1.000**

Rule 4 for **1.0** (3; 1.0)
- if TotalWorkingYears <= 2
- and OverTime = No
- and Age <= 32
- and WorkLifeBalance in [4.000]
- and NJobRole in [3.000]
- then **1.000**

Conclusion

C5 classification methods used to develop a model for predicting recently hired employee attrition. C5 model determined that over time, stock option, age of employee, work life balance, and job involvement are the top five important variables in attrition of recently hired employees. In the previous study, using dataset for employees hired for more than ten years, years at the current role, marital status, distance from home and stock option were the most important variables influencing employee attrition. This indicates that the reasons for employee attrition is different for recently hired and longer hired employees. As the results of analysis shows, the C5 prediction accuracy is much higher for recently hired employees as compared with employees hired more than ten years.

This finding is also 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 types and environments 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