The Connection Between Mentoring Programs and the Employee Engagement

Rowena Ebanks
Northwestern University
RowenaEbanks2016@u.northwestern.edu

Abstract--Every company -- be it a small start-up or a large well-established corporation -- must be composed of certain key elements to be successful. These elements can include: having a vision and business plan that ties into it; having a great product and understanding it through analysis and data; and finding a market niche. The full list is lengthy, but savvy business owners understand that one of the most crucial among these elements is its workforce. People can be a company's greatest asset, and they also have the potential to be a company's biggest liability. Having the right workforce in place is a critical component to a business's success, yet having even one employee who is a poor fit can have a profound impact on business operations in terms of morale, loss of productivity, and the costs associated in dealing with or replacing that employee. In fact, studies reveal that the costs of losing an employee (e.g., recruiting, on boarding, training and productivity) can be anywhere from tens of thousands of dollars to 1.5-2x their annual salary (Bersin, 2013). The purpose of this research is to explore mentoring programs as an aspect of employee satisfaction and engagement -- which in turn ties significantly to employee retention. Utilizing data-mining techniques, discussed in detail further in this paper, I will consider how these programs affect employee satisfaction and engagement, and discuss the factors which constitute the most successful among these programs.

I. Introduction

For years employers have had the luxury of cherry picking human capital from a vast pool of talent. Since the financial meltdown in 2008 and 2009, the number of applicants vs. open positions has been significantly higher. Two 2012 examples support this point: (1) more than one million applicants applied for approximately 2,000 open positions at Proctor & Gamble, and (2) Starbucks received approximately seven million applications for 60,000 open roles. These figures are astounding, but not unusual. A couple of recent major shifts in the workplace, however, mark the reversal of this supply-demand trend in human capital. The first shift is that the US economy is on its way to becoming healthy as evidenced by the housing market recovery, corporations having more money, and the US becoming energy independent -- to name a few examples. The second shift is that, according to some estimates, the baby boomers are retiring at a rate of approximately 10,000 a day (Echols, 2014). Now that the tide of human capital has begun to turn, it is even more important for companies to focus on employee satisfaction and engagement to keep morale and productivity high, and to prevent people from leaving their jobs.

Why do people leave a job? The answers may well be hidden in the reasons an employee remains engaged at work. According to a study by Dale Carnegie Training, there are three main drivers of employee engagement:

➢ relationship with immediate supervisor,

➢ belief in senior leadership, and

➢ pride in working for the company.

It seems the key to engagement for most employees is the personal relationship with their immediate supervisor -- and the supervisor's actions and attitude can enhance engagement or create an atmosphere where the employee becomes disengaged (Dale Carnegie Training, 2012). On the other hand, according to Louis Efron of Forbes.com, there are a couple of main reasons the best employees leave a position:

➢ lack of empathy -- employees do not feel they have a stake in a company that is willing to listen to their needs and cares about their concerns;
➢ no future -- if a company does not create clear career paths for their employees that are both communicated and understood, then an employee has little motivation to stay once they wish to grow. (Efron, 2013.)

The focus of this research is to establish whether mentoring programs influence job satisfaction and engagement, which in turn influences employee retention. The objective of this project is threefold: (1) to determine whether mentoring programs have a positive influence on employee satisfaction and engagement, (2) to determine whether mentoring programs have more of an effect on one (or more) demographic than any others, and (3) to define the program format preferred by those employees with a higher degree of employee satisfaction.

The first two sections of this paper will concentrate on a general overview of the major Data Mining Themes, as well as the Cross Industry Standard Process for Data Mining. Next, I will provide some examples of data mining applications specifically used within workforce analytics. Then we will turn our attention to the research that is the subject of this paper to discuss: the specific data used, the steps taken to prepare it for research, and the data mining algorithms applied to meet the stated objectives. Finally, an in-depth account of the data mining results will be examined, followed by my conclusion and recommendation for future work.

II. Data Mining Themes

(Reproduced from Assignment 1)

Data mining is an integral part of data science in that it is the "analyze" step involving computational techniques and algorithms to discover patterns in data -- thus redefining the data into an understandable structure for additional use. At a high level, the goal of data mining tends to be either descriptive (i.e., extrapolating patterns that describe relationships in data) and/or predictive (i.e., making predictions based on existing data). These goals, or tasks, can be further classified into the following major themes of data mining (Fayyad, Piatetsky-Shapiro, Smyth, 1996).

➢ *Classification* attempts to predict which class new data will fall into by using historical patterns, and generalizing known structures, in data.

➢ *Regression* attempts to predict, or estimate, the value of a variable. Regression is also known as "value estimation" as it seeks to predict *how much* something will happen, e.g., how likely is a customer to click on a promotional link?

➢ *Clustering* attempts to categorize data into groups that are in some way similar, without using known structures in the data.

➢ *Association rule learning*, or dependency modeling, attempts to find relationships between variables. An example of this is market basket analysis where a grocer might gather customer-purchasing data to determine which items are frequently bought together.

➢ *Anomaly detection* attempts to identify unusual data records that might require further investigation. An example of this is fraud (credit card, medical insurance, etc.).

➢ *Sequential pattern mining* attempts to reveal patterns between data examples where the value occurs together frequently in sequence, e.g., stock trend prediction, web user analysis.

➢ *Link prediction* attempts to predict some sort of link between data items, typically by implying there is a link between the two. This is common on social network sites like LinkedIn.com where suggestions are given to you to connect with people you might know based on your existing list of connections.

Note: this list is not comprehensive, but is intended to provide some of the best examples of the algorithms and techniques used to uncover patterns in data.

III.CRISP-DM Methodology

(Reproduced from Assignment 1)

The Cross Industry Standard Process for Data Mining (CRISP-DM) is a process model for data mining which describes commonly used methodologies for approaching and solving data problems. As the figure and discussion illustrate, there are six stages in the CRISP-DM model:

1) Business Understanding,
2) Data Understanding,
3) Data Preparation,
4) Modeling,
5) Evaluation, and
6) Deployment.

Business Understanding is the initial stage of the data mining process, and approaches the problem from a business perspective. The focus in this stage is understanding the project objectives and requirements. The information is then utilized to define the data mining problem and a preliminary plan is outlined to achieve the project objectives. The tasks in this stage involve determining business objectives, assessing the current situation, determining the data mining goals, and producing the project plan. Some might view this as the easiest stage in the process; however, it is extremely crucial to recognize the importance of this stage -- you must know and understand the problem in order to solve it.

The *Data Understanding* stage starts with the initial data collection (what data do we have, and what data do we still need), and follows with activities that will allow us to understand the data, discover problems related to data quality, uncover first insights into the data, and/or locate interesting subsets in order to form hypotheses regarding hidden information. The tasks involved in this stage involve collecting initial data, describing the data, exploring the data, and verifying the data quality.

The *Data Preparation* is the process of transforming the raw data into data that will be fed into the modeling tool(s) -- this transformed data is more commonly referred to as the final dataset. The activities involved in this process (e.g., table, record, and attribute selection; and data transformation and cleaning for modeling tools) are often performed multiple times, and in no particular order. The general tasks in this stage are selecting the data, cleaning the data, constructing the data, integrating the data, and formatting the data. Often this stage takes the most time because the data comes from many sources, in many forms.

The *Modeling* stage is where various applications of modeling techniques are chosen, and their parameters are calibrated to optimal values. There are typically a number of techniques for the same data mining problem type, and some techniques place specific requirements on the data form. Because of this, it is often necessary to return to the previous data preparation phase to ensure the appropriate modeling techniques are chosen. The tasks involved in this stage include selecting the modeling techniques, generating the test design, and building and assessing the models. The iterative nature at this stage is crucial to the data mining process.

The goal at the *Evaluation* stage is to reach a decision on how best to utilize the results of the data mining project. In order to do this, the team must revisit the model (or models) it has built -- which, from a data analysis perspective, appear to have high quality. Before final deployment of the model, we must be certain the model properly achieves the objectives set out in the initial Business Understanding stage. To achieve this certainty, we must thoroughly evaluate the model and review the steps taken in order to create it. Another key objective during this evaluation is to establish that all important business issues have been sufficiently covered. The tasks at this stage involve evaluating the results, reviewing the process, and determining next steps.

The final stage in the data mining process is *Deployment*. It's true, a model has been created and the results of the modeling have given us insights into the problems we want to address, but it's important to remember that the data mining process isn't complete until we act on those results to improve the issues that initiated the process. Sometimes the actions necessary can be as simple as generating a report, or as complex as implementing a repeatable process, or processes, across the enterprise. The tasks involved in this stage involve plan deployment, plan monitoring and maintenance, producing and presenting a final report, and reviewing and documenting the project. (Chapman, et al., 2000.)

IV. Examples of Data Mining Applications in Workforce Analytics

Most companies make workforce decisions (e.g., hiring, promotion and career development) based on unreliable or unscientific methods. In other words, these workforce decisions are based more on gut feelings as opposed to analytics-based methodologies -- but this has begun to change.

➢ Black Hills Company -- a Rapid City, South Dakota, utilities company -- found its workforce almost double to approximately 2,000 employees after a 2008 acquisition. The average age of certain groups of workers jumped to 50 from 45, and forecasting showed that within seven years approximately 25% of its workforce would be up for retirement. Because Black Hills is bound to provide electric power and gas with minimal interruptions by state regulators, this much turnover would pose an enormous challenge. The utilities company turned to big data and HR predictive analytics for help. As a result, three years later they had a formal plan for replacing retiring workers, which in turn has changed the way it recruits, hires and trains. Another change is Black Hills now works with technical schools to create training programs to fill highly-technical positions earlier. Additionally, the company is encouraging workers to stay on the job longer by offering retirement-readiness and financial-planning benefits to employees 50 and older (Rafter, 2013).

➢ For each of its 317,500 employees worldwide, Hewlett-Packard generates a “flight risk” score that predicts which one will quit his or her job. It does this so that managers may preemptively intervene where possible, or otherwise plan accordingly (Siegel, 2013). There were likely several data mining themes HP employed to arrive at these scores. Some of the most obvious would include: clustering to categorize data into similar groups, association to attempt to find relationships between the variables, and regression to estimate the value of the likelihood an employee would leave.

➢ Through analytics, executives at Google have calculated the value of top employees and discovered the performance differential between an average technologist and an exceptional one can be as much as 300 times higher. As a result, Google provides its staff the resources necessary to hire, keep, and develop exceptional talent (Sullivan, 2013). (*Reproduced from Assignment 1)*

➢ According to a Dashburst study, there is a high cost associated with unhappy employees lost productivity costs the country approximately $200 billion a year, and lost work days due to stress costs approximately $30 billion a year. By contrast, happy employees stay at their jobs 2x as long (reducing the costs of recruiting and training), and production increases because sick leave decreases. Satisfied employees are more likely to volunteer in assisting their coworkers, and strongly identify with the company culture (Hughes, 2013). (*Reproduced from Assignment 1)*

Mentoring programs, as the focus of this research, can help facilitate sharing experiences, knowledge, insights, skills, and expertise through collaborative learning and dialogue. Savvy organizations recognize workplace mentoring programs can serve the entire employee lifecycle. According to a Chronus.com whitepaper, mentoring new employees can improve new-hire retention rates, as well as enable a company’s succession plans. Additionally, through reverse mentoring, a company can ensure senior executives stay current with technology and innovative business practices.

V. Data Understanding

The data used in this project was compiled by survey, written specifically for this project, using the Google Consumer Surveys service. The survey is shown across a network of premium online news, reference and entertainment sites where it is embedded directly into content, as well as through the Google mobile app. On the web, people answer questions in exchange for access to that content, an alternative to subscribing or upgrading. Google infers the person’s gender, age, and geographic location based on their browsing history and IP address. On mobile, people answer questions in exchange for credits for books, music, and apps. They answer demographic questions up front which means a representative sample of thousands of respondents is automatically built, eliminating the need to ask demographic questions in the survey (Google Consumer Surveys).
The survey was targeted by screening the first question for either a “Yes, previously” or “Yes, currently”

response. The data is external, nominal, and web/social-based. The following are the full survey questions and answers:

1) Have you ever been mentored in a workplace mentoring program? *(target question)*
a) No, I’ve never been mentored
b) Yes, previously
c) Yes, currently

2) Has being part of a mentoring program affected your employee satisfaction?
a) It has increased my satisfaction
b) It has decreased my satisfaction
c) No, my satisfaction remains high
d) No, my satisfaction remains low

3) Was your mentoring program more casual or formal in nature?
a) Casual, the way I prefer it
b) Casual, but I prefer more formal
c) Formal, the way I prefer it
d) Formal, but I prefer it more casual

4) How often were your mentoring meetings
a) Fewer than once a month
b) Once a month
c) More than once a month

5) Did you like the frequency of your mentoring meetings?
a) Just about right
b) Prefer more often
c) Prefer less often

6) What was the overall length of your mentoring program?
a) 1-6 months
b) 7-12 months
c) over 12 months

7) If you have the opportunity again, would you participate in another mentoring
program?
a) Yes, it’s important to career development
b) I’m indifferent
c) No, once was enough

The list of attributes for the raw data gathered is:

User ID	Parental Status	Question 1 Response Time
Time	Question 1 Answer	Question 2 Response Time
Publisher Category	Question 2 Answer	Question 3 Response Time
Gender	Question 3 Answer	Question 4 Response Time
Age	Question 4 Answer	Question 5 Response Time
Geography	Question 5 Answer	Question 6 Response Time
Urban Density	Question 6 Answer	Question 7 Response Time
Income	Question 7 Answer	

VI. Data Preparation

In the data preparation phase, I eliminated the attributes that were numeric, or that I felt were not relevant to my research. The chart below illustrates which attributes were deleted -- the attributes in **bold** were retained, and the attributes in gray strikethrough
font were deleted as below:

~~User ID~~	~~Parental Status~~	~~Question 1 Response Time~~
~~Time~~	**Question 1 Answer**	~~Question 2 Response Time~~
~~Publisher Category~~	**Question 2 Answer**	~~Question 3 Response Time~~
Gender	**Question 3 Answer**	~~Question 4 Response Time~~
Age	**Question 4 Answer**	~~Question 5 Response Time~~
Geography	**Question 5 Answer**	~~Question 6 Response Time~~
Urban Density	**Question 6 Answer**	~~Question 7 Response Time~~
Income	**Question 7 Answer**	

Additionally, I modified the attribute values where necessary by removing special characters and shortening values. Details of this data transformation can be found in Charts 1-3 of the Index.

VII. Data Mining Algorithms Utilized

The focus of this research is to determine whether mentoring programs influence job satisfaction and engagement, which studies show affect employee retention. Three separate themes were used to meet the three objectives set forth in the Introduction.

In objective 1, the goal was to determine whether mentoring programs have a positive influence on employee satisfaction and engagement. The survey results are clear on this point; in fact, 58% of all respondents replied that their employee satisfaction has increased as a result of being in a mentoring program.

Beyond these results, however, I wanted to examine how accurately I could predict an increase in employee satisfaction. Because I am attempting to predict a value of new data (i.e., predict an increase in employee satisfaction), I employed the classification technique of data mining as I used historical patterns to make the prediction. Specifically, I compared the J48 Decision Tree algorithm to the Naive Bayes algorithm to see which returned a better model.

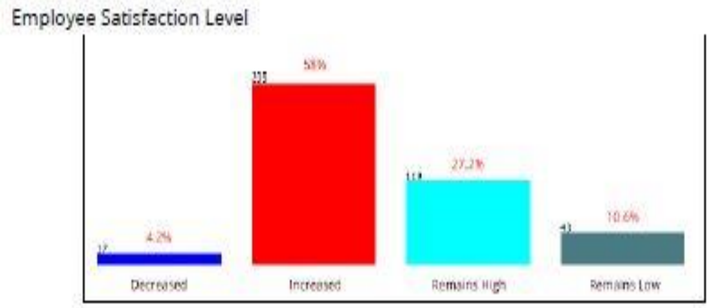
For <u>objective 2</u> I wanted to determine whether mentoring programs have more of an
effect on one demographic than another, thereby developing a profile of the group that
appears to benefit most from mentoring programs. The theme used for this objective
was clustering as the goal was to categorize data into groups that are in some way similar. I used both the DBScan and kMeans algorithms to compare the clusters returned. Ultimately I chose to focus solely on the kMeans algorithm as the clusters returned with DBScan were all categorized as noise regardless of the parameters entered.

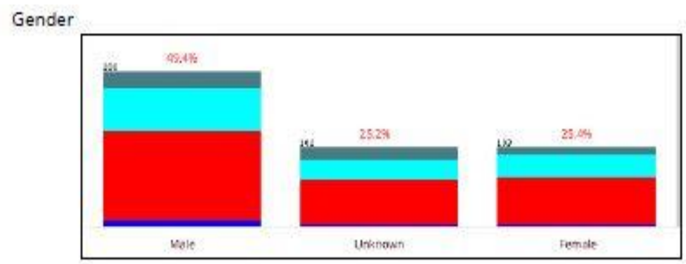
Finally, for <u>objective 3</u> I wanted to define the program format preferred by those people with a higher degree of employee satisfaction. Because I wanted to find relationships between variable, I chose the dependency modeling technique known as association rule learning. The specific algorithm used for this theme was Apriori for its generate-and-test approach (i.e., it generates candidate item sets and tests if they are frequent).

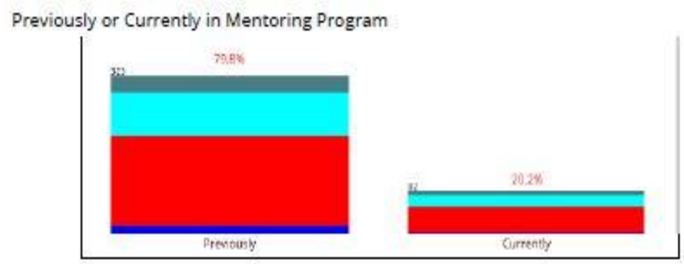
VIII. Experimental Results and Analysis

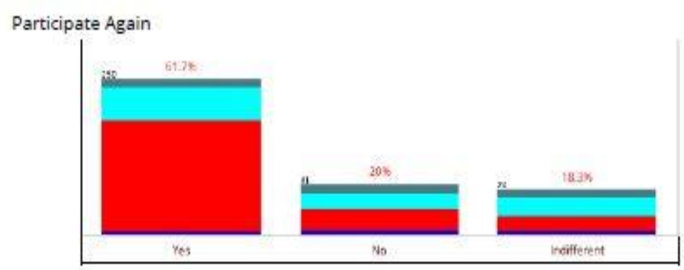
Before presenting the results and analysis, I would like to present the overall results of the survey. In total there were 5,194 responses, of which 405 (7.8%) are currently, or were previously, in a mentoring program. Of the 405 targeted respondents, gender was known for 303 respondents and unknown for 102 respondents.

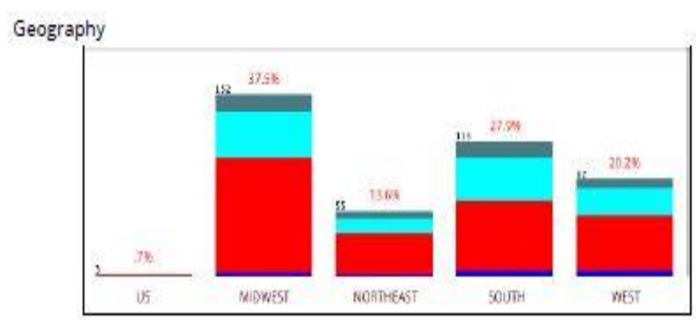
The following five charts present the most relevant overall findings: employee satisfaction level, gender, currently/previously in a program, geography, and whether the respondent would participate in a program again. A quick overview shows that 58% of respondents had an increased employee satisfaction, only 20% were in a program at the time of response, but 62% would participate again because they believe it is important to career development.









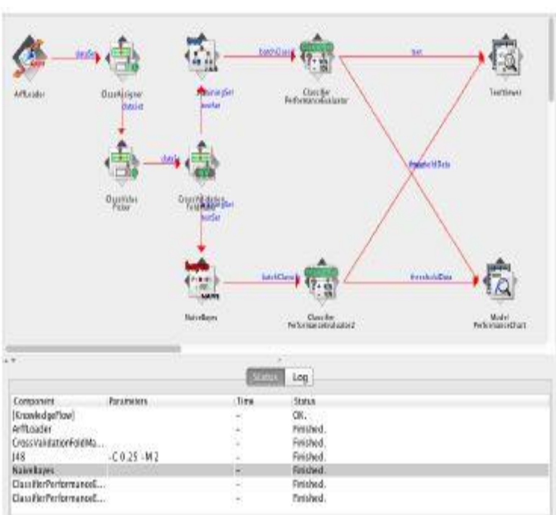


The tool used for this analysis was WEKA, developed at the University of Waikato, New Zealand. It is a popular suite of machine learning and data mining software written in Java, and as such is a powerful resource to help businesses better understand and improve performance through analytics.

OBJECTIVE 1: Predict whether there will be an increase in employee satisfaction.

ALGORITHM: a comparison of J48 Decision Tree and Naive Bayes
DATASET (303 instances): targeted known gender only (i.e., male/female) dataset
to predict increase in employee satisfaction
ATTRIBUTES (10): Gender, Age, Income, Currently/Previously in Program, *EE Satisfaction Level*, Formal/Casual, number of Monthly Meetings, Meeting Frequency Preference, Length of Program, and Participate Again

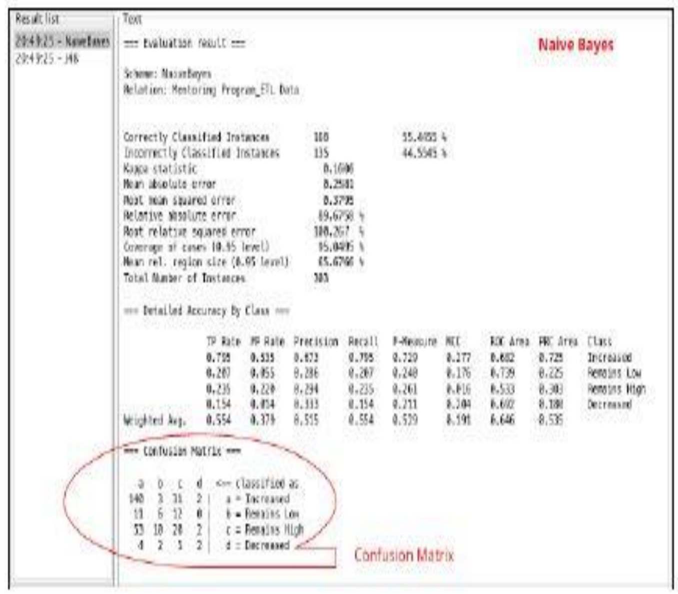
To compare the two algorithms, I used the KnowledgeFlow application in WEKA, and created the outline below. Note that WEKA defaulted to "Participate Again" as a target variable. In order to control the target, I incorporated the ClassValuePicker to select "EE Satisfaction Level." The following illustration shows the WEKA process flow.



The results of the algorithms are outlined in the following two illustrations.

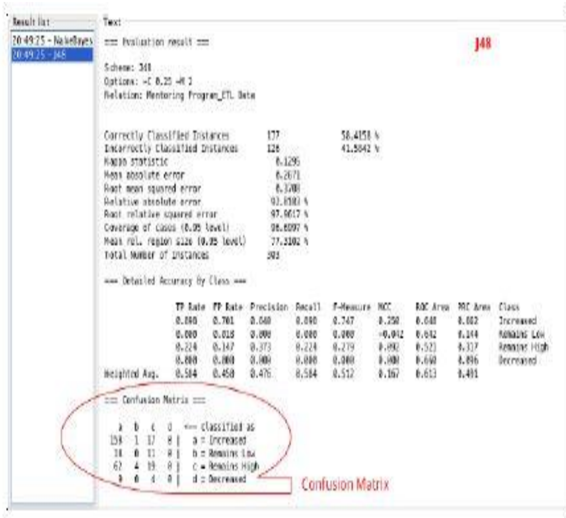
Results of Naive Bayes

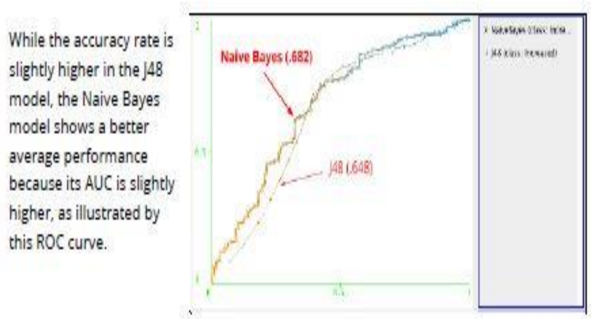
➔Accuracy 55.45% (error rate is 44.55%)
➔ TP (actual "increased" correctly classified as "increased") is 140
➔FP (actual non"
increased" incorrectly labeled as "increased) is 68
➔TN (actual non"
increased" correctly classified as non"
increased") is 28
➔FN (actual "increased" incorrectly labeled as non"
increased") 67
➔AUC for our first model is .682



Results of J48

➔Accuracy 58.42% (error rate is 41.58%)
➔ TP (actual "increased" correctly classified as "increased") is 158
➔FP (actual non"
increased" incorrectly labeled as "increased) is 89
➔TN (actual non"
increased" correctly classified as non"
increased") is 19
➔FN (actual "increased" incorrectly labeled as non"
increased") 37
➔AUC for our first model is .648





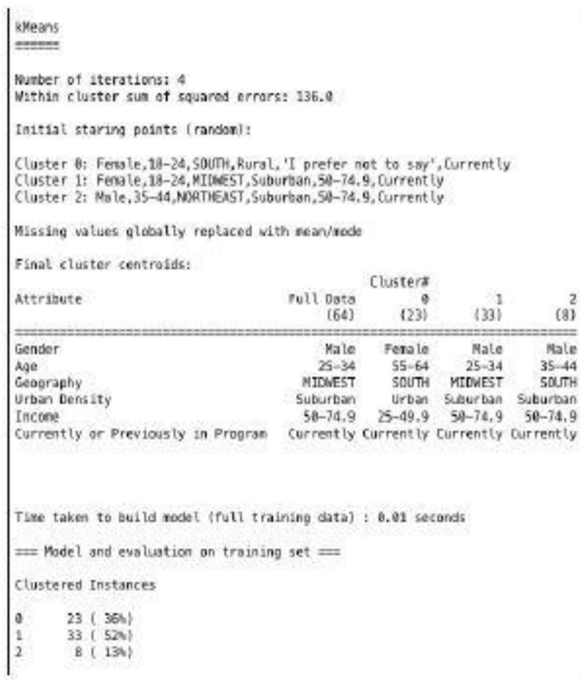
OBJECTIVE 2: Determine demographic/profile most-positively affected by mentoring programs.

I chose to examine respondents currently and previously in programs separately to see if I could glean any differences between the two groups. The results are outlined below.

ALGORITHM: kMeans
DATASET (64instances): targeted known gender with EE Satisfaction Level at *Increased* or *Remains High* where respondents were *Currently* in a program
ATTRIBUTES (5): Gender, Age, Geography, Urban Density, Income

Results of kMeans on *Currently* in program respondents

➔Cluster descriptions (number of clusters = 3)
● Full dataset
○ Male, 25-34, MIDWEST, Suburban, $50,000-$74,999
○ 64 instances -- 100%
● Cluster 0
○ Female, 55-64, SOUTH, Urban, $25,000-$49,999
○ 23 instances -- 36%
● Cluster 1
○ Male, 25-34, MIDWEST, Suburban, $50,000-$74,999
○ 33 instances -- 52%
● Cluster 2
○ Male, 35-44, SOUTH, Suburban, $50,000-$74,999
○ 8 instances -- 13%



ALGORITHM: kMeans
DATASET (213 instances): targeted known gender with EE Satisfaction Level at *Increased* or *Remains High* where respondents were *Previously* in a program
ATTRIBUTES (5): Gender, Age, Geography, Urban Density, Income

Results of kMeans on *Previously* in program respondents

➔Cluster descriptions (number of clusters = 3)
● Full dataset
○ Male, 55-64, MIDWEST, Suburban, $25,000-$49,999
○ 213 instances -- 100%
● Cluster 0
○ Female, 55-64, SOUTH, Suburban, $25,000-$49,999
○ 94 instances -- 44%
● Cluster 1

○ Male, 55-64, WEST, Urban, $50,000-$74,999
○ 59 instances -- 28%
● <u>Cluster 2</u>
○ Male, 65+, MIDWEST, Suburban, $25,000-$49,999
○ 60 instances -- 28%

```
kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 434.0

Initial staring points (random):

Cluster 0: Female,25-34,SOUTH,Suburban,75-99.9,Previously
Cluster 1: Male,55-64,US,Unknown,Unknown,Previously
Cluster 2: Male,65+,MIDWEST,Suburban,25-49.9,Previously

Missing values globally replaced with mean/mode

Final cluster centroids:
                                                    Cluster#
Attribute                            Full Data          0          1          2
                                         (213)       (94)       (59)       (60)
===============================================================================
Gender                                    Male     Female       Male       Male
Age                                      55-64      55-64      55-64        65+
Geography                              MIDWEST      SOUTH       WEST    MIDWEST
Urban Density                         Suburban   Suburban      Urban   Suburban
Income                                 25-49.9    25-49.9    50-74.9    25-49.9
Currently or Previously in Program  Previously Previously Previously Previously



Time taken to build model (full training data) : 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       94 ( 44%)
1       59 ( 28%)
2       60 ( 28%)
```

MOST-NEGATIVELY AFFECTED CLUSTERING
It struck me while running the data through kMeans for most-positive clustering, that it would be interesting to see if there was anything to be gained from finding the most-negatively affected clustering as well.

<u>ALGORITHM</u>: kMeans
<u>DATASET (6 instances)</u>: targeted known gender with EE Satisfaction Level at *Decreased* or *Remains Low* where respondents were *Currently* in a program
<u>ATTRIBUTES (5)</u>: Gender, Age, Geography, Urban Density, Income

Results of kMeans of respondents *Currently* in program

➔Cluster descriptions (num of clusters = 2)
● <u>Full dataset</u>
○ Female, 55-64, SOUTH, Suburban, $25,000-$49,999
○ 6 instances -- 100%
● <u>Cluster 0</u>
○ Female, 18-24, SOUTH, Rural, $50,000-$74,999
○ 3 instances -- 50%

● <u>Cluster 1</u>
○ Female, 55-64, MIDWEST, Suburban, $25,000-$49,999
○ 3 instances -- 50%

```
kMeans
======

Number of iterations: 2
Within cluster sum of squared errors: 9.0

Initial staring points (random):

Cluster 0: Female,18-24,SOUTH,Rural,'I prefer not to say'
Cluster 1: Female,55-64,MIDWEST,Suburban,25-49.9

Missing values globally replaced with mean/mode

Final cluster centroids:
                                  Cluster#
Attribute            Full Data          0          1
                           (6)        (3)        (3)
====================================================
Gender                  Female     Female     Female
Age                      55-64      18-24      55-64
Geography                SOUTH      SOUTH    MIDWEST
Urban Density         Suburban      Rural   Suburban
Income                 25-49.9    50-74.9    25-49.9



Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       3 ( 50%)
1       3 ( 50%)
```

<u>ALGORITHM</u>: kMeans
<u>DATASET (36 instances)</u>: targeted known gender with EE Satisfaction Level at *Decreased* or *Remains Low* where respondents were *Previously* in a program
<u>ATTRIBUTES (5)</u>: Gender, Age, Geography, Urban Density, Income

Results of kMeans of respondents *Previously* in program

➔Cluster descriptions (number of clusters = 2)
● <u>Full dataset</u>
○ Male, 55-64, SOUTH, Urban, $25,000-$49,999
○ 36 instances -- 100%
● <u>Cluster 0</u>
○ Male, 55-64, SOUTH, Urban, 25,000-$49,999
○ 22 instances -- 61%
● <u>Cluster 1</u>
○ Male, 55-64, SOUTH, Suburban, 25,000-$49,999
○ 14 instances -- 39%

```
kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 77.0

Initial staring points (random):

Cluster 0: Male,55-64,NORTHEAST,Urban,75-99.9
Cluster 1: Male,55-64,SOUTH,Suburban,50-74.9

Missing values globally replaced with mean/mode

Final cluster centroids:
                                      Cluster#
Attribute              Full Data         0           1
                          (36)         (22)        (14)
=========================================================
Gender                    Male         Male        Male
Age                      55-64        55-64       55-64
Geography                SOUTH        SOUTH       SOUTH
Urban Density            Urban        Urban    Suburban
Income                 25-49.9      25-49.9     25-49.9




Time taken to build model (full training data) : 0 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       22 ( 61%)
1       14 ( 39%)
```

OBJECTIVE 3: Define the preferred format of mentoring program by those with a higher degree of employee satisfaction.

To determine the format, I ran the dataset of respondents with an increased/remains high EE satisfaction level.

<u>ALGORITHM</u>: Apriori, minsup=.1
<u>DATASET (345 instances)</u>: targeted known gender with EE Satisfaction Level at *Increased* or *Remains High*
<u>ATTRIBUTES (7)</u>: Previous/Current, EE Satisfaction (increased/remains high), Formal/Casual, Monthly Meetings, Meeting Frequency, Length of Program, Participate Again

	Conf	Lift	LHS				RHS
1)	1	1.26	Monthly Mtgs =Less than one	Length of Program =1-6 months			36 ==> Currently or Previously in Program =Previously 36
2)	0.95	1.2	Formal or Casual =Casual	Monthly Mtgs =Less than one	Mtg Frequency Preference =Just right		41 ==> Currently or Previously in Program =Previously 39
3)	0.93	1.18	Formal or Casual =Casual	Monthly Mtgs =Less than one			57 ==> Currently or Previously in Program =Previously
4)	0.93	1.58	EE Satisfaction Level =Increased	Mtg Frequency Preference =Just right	Length of Program =7-12 months		41 ==> Participate Again =Yes 38
5)	0.92	1.17	EE Satisfaction Level =Increased	Monthly Mtgs =Less than one	Mtg Frequency Preference =Just right		39 ==> Currently or Previously in Program =Previously 36
6)	0.91	1.33	Monthly Mtgs =More than one	Mtg Frequency Preference =Just right	Length of Program =1-6 months	Participate Again=Yes	41 ==> EE Satisfaction Level =Increased 40
7)	0.91	1.16	Monthly Mtgs =Less than one	Mtg Frequency Preference =Just right			65 ==> Currently or Previously in Program =Previously 59

As with the clustering exercise in objective 2, I wanted to see if I could glean any associations that were common among people whose satisfaction level was lower, so I similarly ran the decreased/remains low dataset through the Apriori algorithm.

<u>ALGORITHM</u>: Apriori, minsup=.1
DATASET: targeted known gender with EE Satisfaction Level at *Decreased* or *Remains Low* (60 instances)
<u>ATTRIBUTES (7):</u> Previous/Current, EE Satisfaction (increased/remains high), Formal/Casual, Monthly Meetings, Meeting Frequency, Length of Program, Participate Again

	Conf	Lift	LHS			RHS
1)	1	1.4	Monthly Mtgs =Less than one	Length of Program =12+ months		12 ==> EE Satisfaction Level =Remains Low 12
2)	1	1.2	EE Satisfaction Level =Decreased	Monthly Mtgs =More than one		10 ==> Currently or Previously in Program =Previously 10
3)	1	1.4	Formal or Casual =Casual	Monthly Mtgs =Less than one		10 ==> EE Satisfaction Level =Remains Low 10
4)	1	1.2	Formal or Casual =Casual	Length of Program =1-6 months		9 ==> Currently or Previously in Program =Previously 9
5)	1	1.2	Monthly Mtgs =Less than one	Participate Again =Yes		9 ==> Currently or Previously in Program =Previously 9
6)	1	1.2	Monthly Mtgs =More than one	Mtg Frequency Preference =Just right		9 ==> Currently or Previously in Program =Previously 9
7)	1	1.2	Length of Program =1-6 months	Participate Again =No		9 ==> Currently or Previously in Program =Previously 9
8)	1	1.4	Currently or Previously in Program =Previously	Monthly Mtgs =Less than one	Length of Program =12+ months	9 ==> EE Satisfaction Level =Remains Low 9
9)	0.95	1.14	Length of Program =1-6 months			21 ==> Currently or Previously in Program =Previously 20
10)	0.94	1.13	EE Satisfaction Level =Decreased			17 ==> Currently or Previously in Program =Previously 16

IX. Summary and Conclusion

Objective 1: Summary and Conclusion

In the first objective I wanted to establish whether participation in a mentoring program increased employee satisfaction. From my research, it is clear that one of the benefits of participating in a mentoring program is, indeed, an increase in employee satisfaction. In fact, 58% of respondents said it increased, while 27% answered it remained high as a result of the program. Additionally, I wanted to establish whether I could predict an increase in employee satisfaction with the data gathered. The accuracy of both the J48 and Naive Bayes models was low at 58% and 55% respectively.

In conclusion, while the prediction models were not as accurate as I had hoped, the data still supports implementing a mentoring program in spite of low ability to predict those whose satisfaction will increase. However, if the need exists to predict an increase of satisfaction, I recommend utilizing the Naive Bayes model for its better average performance.

Objective 2: Summary and Conclusion

In the second objective I wanted to profile the group that is most-positively affected by a mentoring program. I created clusters utilizing two separate datasets -- currently in a program, and previously in a program. Additionally, I wanted to see if I could find any additional patterns by clustering those most-negatively affected, utilizing the same categorizations. I chose to focus on the clusters of the full datasets:

<u>Increased/High (using full dataset)</u>
currently: Male, 25-34, MIDWEST, Suburban, $50,000-$74,999
previous: Male, 55-64, MIDWEST, Suburban, $25,000-$49,999

<u>Decreased/Low (using full dataset)</u>
currently: Female, 55-64, SOUTH, Suburban, $25,000-$49,999
previous: Male, 55-64, SOUTH, Urban, $25,000-$49,999

It is interesting to note that those most-positively affected were Males in the Midwest, and those most-negatively affected were between the ages of 55-64 in the South making between $25,000-$50,000. In conclusion, I would recommend focusing on creating more mentoring programs in inner-cities and suburbs in the South. To implement successful programs, and improve upon those already in place, I further recommend examining the formats of the more successful mentoring programs in place in the Midwest.

Objective 3: Summary and Conclusion

For the final objective I wanted to define a general format that was most preferred by those respondents whose satisfaction increased or remained high.

Among those whose satisfaction increased or remained high, the preference was for a more casual program with an overall length of 6 months or less. Many of the association rules included "just right" for the frequency of meetings attribute, regardless of what number of monthly meetings might have been associated with it. What this suggests to me is the frequency of monthly meetings wasn't so much an issue -- rather, respondents seemed satisfied to just be in a mentoring program. I, therefore, recommend a more casual, flexible, shorter program as the best format from these associations.

Among those with a satisfaction level that decreased or remained low, there wasn't much information I pulled from the association rules except for one. In Association Rule 5, one of the variables was that respondents would participate again given the opportunity. What this suggests is that even though respondents' satisfaction levels remained low, they still believe mentoring programs are an important aspect of career development.

X. Future Work

To continue the work on this project, the next steps are clear -- mentoring programs can be an important aspect of employee satisfaction and it would be wise to implement them as a general practice. The good news is that the programs don't need to be formal to have impact. A shorter, more casual and flexible program is the most preferred and successful. Research into programs implemented in the Midwest may help to refine the format a bit, and more of a focus should be targeted in the South.

Beyond implementing more mentoring programs, it would be wise to begin combining business data (sales, revenue, output, margin, quality, etc.) with the mentoring data for more compelling findings. Businesses can now begin to gain true insights into the correlation between business performance and their workforce. These insights will likely have significant impacts on customer satisfaction as well as operating metrics. A strong partnership between HR and the business is vital, and the insights gained through this type of analytics can help determine important actions that can yield considerable profit and loss impact, and drive increased business performance.

References

Bersin, J. (2013, August 16). Employee retention now a big issue: Why the tide has turned.
Retrieved from
https://www.linkedin.com/today/post/article/20130816200159-131079-employee-retention-now-a-big-issue-why-the-tide-has-turned
Chapman, P., Clinton, J., Kerber, R., Khabaza, T. Reinartz, T., Shearer, Co., Wirth, R. (2000).
CRISP-DM 1.0: Step-by-step data mining guide. SPSS, Inc.

Dale Carnegie Training (2012). What drives employee engagement and why it matters
[White paper]. Retrieved from
http://www.dalecarnegie.com/assets/1/7/driveengagement_101612_wp.pdf
Echols, M. E. (2014, April 4). Why CLOs should care about retention. Chief Learning
Officer. Retrieved from
http://www.clomedia.com/articles/why-clos-should-care-about-retention
Efron, L. (2013, June 24). Six Reasons Your Best Employees Quit You. Retrieved from
http://www.forbes.com/sites.louisefron/2013/06/24/six-reasons-your-best-employ
ees-quit-you/
Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996). From Data Mining to Knowledge
Discovery in Databases. AI Magazine. Fall 1996. Retrieved from
http://www.csd.uwo.ca/faculty/ling/cs435/fayyad.pdf
Google Consumer Surveys. http://www.google.com/insights/consumersurveys/home.
Retrieved from
http://www.google.com/insights/consumersurveys/how
Hughes, S. (2013, June 24). The high cost of unhappy employees to businesses
[infographic]. Dashburst. Retrieved from
http://dashburst.com/infographic/unhappy-employees-business-cost/
Siegel, E. (2013). *Predictive Analytics: The Power to Predict who will Click, Buy, Lie or Die*.
Hoboken, NJ: John Wiley & Sons, Inc.
Sullivan, J. (2013) How Google became the #3 most valuable firm by using people
analytics to reinvent HR. *ere.net*. Retreived from
http://www.ere.net/2013/02/25/how-google-became-the-3-most-valuable-firm-byusing-
people-analytics-to-reinvent-hr/

INDEX

Chart 1

Chart 2

Chart 3

The Google Doc link for this Final Research Project can be found at
(https://docs.google.com/a/u.northwestern.edu/document/d/10VXTlitMQVkX14n788O4re
V4My0HzNnL4k9cxoPNxTc/edit#)
All datasets in Google Docs can be found at Mentoring Program -- DATA link
(https://docs.google.com/a/u.northwestern.edu/spreadsheets/d/1BiwQhGtSGsOIEn8Qrl4
dBHhDtDUfSSYM37MOaRK_dyA/edit#gid=0)