CSC-272: Final Project

Allison Fierce, Sarah Gauby, and Laura Peckham

1 Introduction

Who gets the job? Its a question many wish they knew the answer to. Using data from a survey of recent college graduates of various ages and levels of education, we searched for the answer to this question by working to find the best predictive models. The first question we addressed was how to predict whether a given individual was in the labor force. We attempted to predict labor force status using several different variations of our dataset. After a thorough review of the potential models on these datasets, we found that the best predictive models were produced by J48 and IBK. In addition to labor force status, we looked at a variety of other class attributes and decided to find the best predictive models for gender, salary, age, and job satisfaction. For these, J48 and IBK once again proved to be the best models for making the desired prediction.

2 Dataset Description

The data we analyzed came from a survey of college graduates, regarding their occupational status, and related factors, such as level of education and field of work. We obtained our data through the National Science Foundation (NSF) website. From the main page, we went to NCSES Data, under the Statistics Menu. From there we clicked on Public Use Files, and then on to Natinal Survey of College Graduates, which brought us to the following page: http://sestat.nsf.gov/datadownload/. From there we downloaded the 2010 file titled SESTAT (Integrated Survey Data). Our data contained 108337 instances, and 139 attributes. For a description of the attributes we examined, see the appendix.

3 Data Preparation

Because we had so many attributes, we wanted to narrow them down. We went through our data in Excel, deleting attributes we felt were redundant or irrelevant, like the reference year, as that was 2010 for every entry. We also deleted attributes that only applied to a small percentage of respondents, such as questions geared toward those receiving government support, and

attributes that were subsets of other attributes, such as questions about job satisfaction. While we kept the data for overall satisfaction, we did not keep the attributes for satisfaction in specific areas, including geographic location, opportunity for promotion,and salary. Total, we deleted 90 attributes, leaving 49 remaining. At this stage, we replaced every response of L, or Logical Skip, with ?, which denotes missing data in Weka. We then converted our data into arff format. We reviewed our data to make sure it was correct, taking extra care that each attribute was in the appropriate form. Many of our nominal attribute-values were designated numbers, leading them to be falsely labeled as numeric. Once this was complete, we created several different data sets, each suited to different analyses. While most of our data was nominal, we were also interested in numeric estimation of the numeric attributes. We created one dataset where we discretized all numeric attributes (into 10 bins), in order to be able to run Apriori (association learning), and left the other with numeric data. We next addressed missing values. We kept the data with missing values, as missing values could still reveal interesting connections in the data. We noticed that we had a group of 18,373 respondents consistently skipping questions, so we created another set (both with the numeric and discretized original data) that removed these instances. We created a third set from the discretized data where we removed any attributes with missing data. The 15 remaining attributes had no missing values. For each of these sets, we shuffled up the data, and divided it into two sets: one containing eighty percent of the instances, to build models with,and one containing twenty percent of the instances, to test our final models on. Finally, we created a different dataset containing only the attributes related to job satisfaction, to run separate tests on. We will refer to our datasets as the following:

Dataset 1: Complete data including missing values (discretized and numeric)
Dataset 2: Dataset with any instances that contain missing values removed (discretized and numeric)
Dataset 3: Dataset with any attributes that contain missing values removed
Dataset 4: Job satisfaction data (discretized and numeric)

With our datasets prepared, we were ready to start analyzing them.

4 Data Analysis

During our analysis, we utilized the following basic algorithms:

OneR: Uses a single attribute to predict the outcome.
J48: Creates a decision tree.

Naive Bayes: Regards all attributes as independent and equally important, whether they are or not. Uses conditional probabilities derived from the training data, to calculate the likelihood of an outcome being a particular class.
K-Nearest Neighbor (IBK): Compares a new instance to all existing instances and classifies it according to how the K most similar instances are classified.
Apriori: A form of association learning. Starts with 1-itemsets and gradually builds larger sets, based on frequency of the sets within the data. Then creates rules from each itemset, keeping the rules above a certain accuracy.
Linear Regression: Builds a linear model based on the attributes Weka deems relevant. Works best when attributes are independent.

Dataset 1 Analysis: We first looked at OneR, to get a sense of the original data, including missing attributes. We next looked at association rules, to get an idea of the relationships within the data. Upon looking at the association rules, we noticed that many of the rules showed redundant attributes. For example, one rule told us that an individual born in the United States would be a citizen of the United States. This led to more tests being used in the experiment after the removal of some redundant data. Redundant attributes would have no effect on OneR, but they do affect Naive Bayes and K-nearest neighbor, or IBK. After removal of the redundant attributes, we ran Naive Bayes and IBK with a K of 5, 10, and 20. We then ran J48 to create a tree, using the data with the removed redundant attributes. We stipulated that each leaf of the tree should contain at least 100 instances, given the volume of our data.

Dataset 2 Analysis: We first looked at Dataset 2 without discretizing it. We did not feel that this analysis added to our understanding of the data, nor did it produce any models or rules with a better accuracy than the discretized dataset, so we decided to focus on the discretized portion of this dataset. For the discretized data set which has instances with missing attributes removed, we initially researched the percentage of correctly classified instances in the algorithms OneR (as a baseline), IBK with a K of 5, J48 with the default confidence of .25 and instances per leaf (m) of 2, and Nave Bayes for the class attributes LFSTAT (Job Satisfaction), Salary, Age, and Gender. These attributes were chosen because the ability to predict these would be useful in a real-world situation. The results of these tests allowed us to determine that Gender was the most accurately predicted class attribute without any alterations to make the algorithms more specific to the attribute, and Salary was the least accurately predicted. We determined that focusing on the best models for these two attributes would best allow us to understand the nature of the data. We would use the worst predicted to understand what

aspects could be altered to produce a better data set and the best predicted to construct a meaningful model that could be applied to future instances. To improve the models, we found the best K value for IBK and the optimum combination of confidence and m value for J48 that produced the most correctly classified attributes. We chose to depart from the standard K=5, 10, and 20 for this dataset because since we focused on the discretized data, we wanted to go more in depth with it. For Salary, we looked at a range of 40 K values between 5 and 300 and determined the optimal K was 93. After reviewing a range of m values between 1 and 100 and confidences between .15 and .50, the optimal values were m of 35, confidence of .20. For Gender, we looked at 40 different K values ranging between 1 and 71 to determine a K of 61 was optimal. After reviewing an m value range of 2 to 50 and a confidence range of .23 to .35 we determined the optimal values were an m of 19 and a confidence of .29.

Dataset 3 Analysis: We used Dataset 3 to predict an individuals labor force status. After removing any attributes with missing values, there were 15 attributes remaining, including the class attribute of LFSTAT. We first ran OneR as a baseline for accuracy. We then tried J48, with the condition that each leaf of the tree created had to contain at least 50 instances. We ran IBK with 5, 10, and 20 neighbors, and ran Naive Bayes as well. Lastly, we looked at Apriori for association rules. As with Dataset 1, we noticed that the association rules linked dependent attributes together to give us trivial rules. We even saw the same rule linking birth in the United States with citizenship of the United States. As we know both Naive Bayes and IBK are affected by redundant and dependent attributes, we went through our data, removing these attributes to the best of our abilities. We removed attributes that overlapped. For example, since the answers for HIGHDEGREE (highest degree obtained), and MRDEGREE (most recent degree obtained) were often the same, we chose to only keep HIGHDEGREE. On this edited dataset, we ran Naive Bayes and IBK once more.

Dataset 4 Analysis: Dataset 4 deals with job satisfaction. The attributes tell us satisfaction with various aspects of a job, such as salary, potential for advancement, and location. Each one is ranked on a scale of 1-4, with 1 being very satisfied, and 4 being very dissatisfied. We first treated the attributes as numeric and ran numeric estimation. We then designated them as nominal values, and ran OneR, J48 (with 40 instances per leaf), IBK (with 5, 10, and 20 neighbors), Naive Bayes, and Apriori. We explored running these tests with various subsets of the attributes, chosen based on highest accuracy, but there was no significant difference between these and the original Dataset 4.

5 Results

Dataset 1 Results: The analysis of the original data, including missing values, revealed several different aspects of the data. OneR created a rule based on the attribute ACTCAP, which denotes whether or not the main activity of a persons work involves computer applications. It correctly classified 96.1567% of the values. However, the rule does not necessarily indicate a causality between working in computer applications, and having a job. This is because whether a respondent answered yes and no, that person was classified as in the labor force. OneR dealt well with the missing values. They were correctly classified as not in the labor force, meaning those who neither have a job, nor are looking for one. We next ran Apriori. Something interesting to note is that although ACTCAP is the rule given in OneR, it is the third association rule, connecting those who do not working in computer applications, to being in the labor force. The first association rule said that those who answered no to having a new business were are likely to be in the labor force. While we would not have predicted this rule, it is not counter-intuitive. Weeks worked is the second rule, which makes sense. Anyone who works 3.7 weeks or more a year is in the labor force by definition. While we might have expected the cutoff to be higher, this is an interesting connection Weka has made, and a viable one. Next, we moved onto Nave Bayes. We obtained an 87.16% accuracy before removing redundant attributes, and an accuracy of 87.2% after. It is an increase in accuracy, but barely. Before removing redundant attributes, our best IBK accuracy is 87.78%, with K=5. After removing redundant attributes, this is bumped up to 87.97%. It is again a miniscule difference that does not seem significant. When we ran the J48 algorithm on the data with the removed redundant attributes and split into nodes with a minimum number of instances of 100, it had an 86.3794% accuracy. The first split was whether or not the person participated in Training. The next split was by involvement in relevant clubs/networking groups, and then by age, gender, and the year of the highest degree. A comparison of the performance of the algorithms we looked at can be seen in Figure 1. We also ran linear regression on this data to predict salary. The correlation coefficient was not as high as we had hoped, but when we visualized our data we saw why. There was a general linear increase in salary as age increased at first, but once the model reached a certain age, there was a decline in salary as people neared retirement. This relationship can be seen in Figure 2. We then visualized weeks worked versus age, and saw a similar trend, as seen in Figure 3. The y-axis on this graph corresponds to ranges of weeks. A y-value of 1, 2, 3, or 4 corresponds to 1-10 weeks, 11-20 weeks, 21-39 weeks, or 40-52 weeks respectively. Figure 3 reinforces the trend we saw in Figure 2.

Dataset 1 Results

Class Attribute: LFSTAT		
Algorithm	**Training Accuracy**	**Holdout Accuracy**
OneR	96.1567%	95.9478 %
J48	86.3563%	86.6479 %
Naive Bayes	81.5664%	87.031 %
IBK (5 neighbors)	87.7812%	87.9033 %
IBK (10 neighbors)	87.4813%	87.6817 %
IBK (20 neighbors)	87.1986%	87.4371 %

Class Attribute: LFSTAT		
Algorithm	**Original Data**	**With Redundant Attributes Removed**
Naive Bayes	87.031 %	87.1986%
IBK (5 neighbors)	87.9033 %	87.9716 %
IBK (10 neighbors)	87.6817 %	87.8274 %
IBK (20 neighbors)	87.4371 %	87.6889 %

Figure 1: Accuracy on Dataset 1

Dataset 2 Results: When we used OneR to predict Salary, we obtained an initial accuracy of 21%. As IBK and J48 performed the best, we manipulated these models to try to improve this percentage. We did the same for Gender. IBK proved to be the optimal model for predicting both attributes. The IBK model constructed for Salary, with a K of 93, had 32.785% correctly classified attributes. This was only 3% improved from the default IBK model, so there was not much we could do. We were not sure why salary was so unpredictable. Perhaps persons not in the job market affected our results, as their salaries are recorded as zero dollars. Gender was the attribute that gave us the greatest OneR accuracy, 66.6%. While we were aware that there are gender differences between majors, we thought the main difference was between the sciences and the humanities. Since this survey was only for sciences majors (but included social/soft sciences), we werent expecting Gender to give us the most accurate predictions. We found it to be a very interesting result, and would enjoy seeing a model that could be used in the future to predict the gender of a recent graduate given the information released in these surveys. After experimenting in Weka, the results for the Gender class attribute showed that an IBK model with a K of 61 correctly classified 70.6108% of its instances. This was an improvement of 1.4843% from the default IBK settings. The complete table of accuracies can be seen in Figure 4

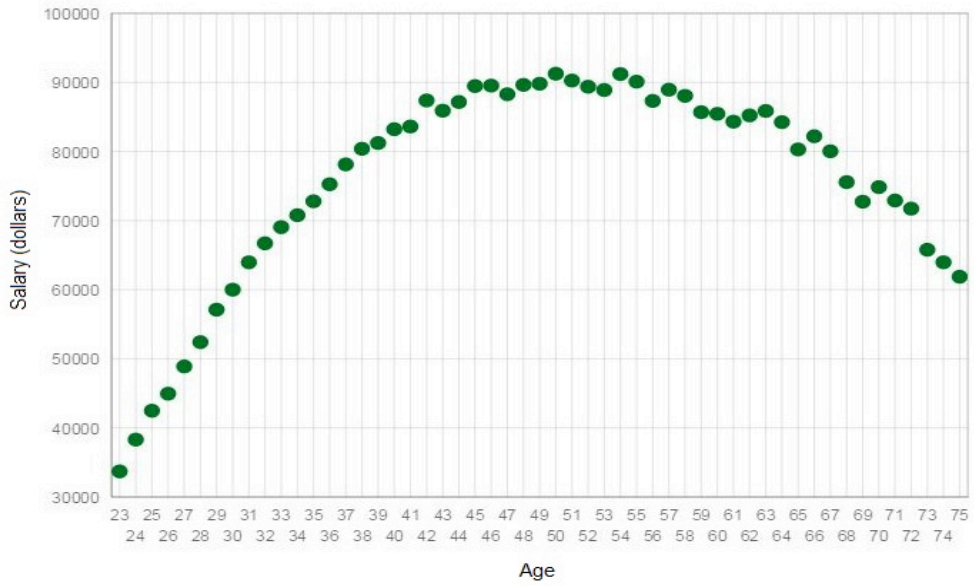

Figure 2: Salary vs. Age

Dataset 3 Results: Figure 5 shows the accuracies of the various models on Dataset 3. J48 gave the highest accuracy, at 85%, then OneR, then IBK with 10 neighbors, and redundant attributes removed. Remember that we are trying to predict a persons status within the labor force. There are three different options: 1=employed, 2=unemployed, and 3=not in the labor force. OneR gave us the following rule:

$$\text{AGE: } \begin{array}{l} \leq 69.8 \longrightarrow 1 \\ > 69.8 \longrightarrow 3 \end{array}$$

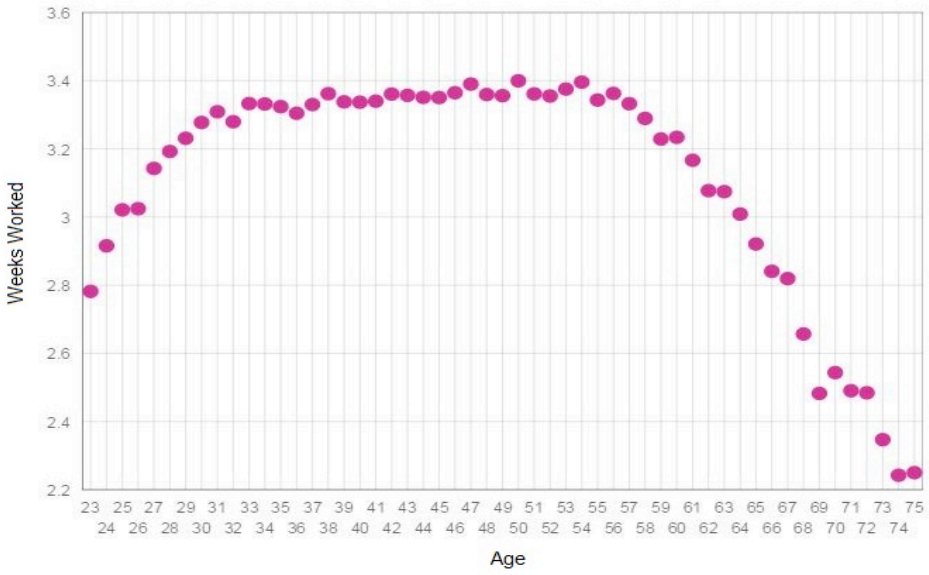

Figure 3: Salary vs. Age

It makes sense that age determines employment, as it is typical to work until retirement, and retirement is determined by age. Because age was the attribute chosen for OneR, we can draw the conclusion that when an individual is not in the labor force, it is most likely due to retirement, rather than maternity/paternity leave, or another reason. We believe this rule will generalize well, as the age of retirement is fairly consistent across the board. Note that class 2, unemployed, is not represented in this rule. A challenge we faced in our prediction was unequal numbers in each group. Most respondents–in fact 83%, were employed, making it easier to predict that class. With the age breakdown, our models also managed to predict class 3 well enough, but not class 2, which accounts for just 4% of instances. In fact, only two of

Dataset 2 Results

Class Attribute: SALARY		
Algorithm	**Training Accuracy**	**Holdout Accuracy**
OneR	21.039%	
J48 (conf. .2, m 35)	31.1014%	31.0121%
Naive Bayes	29.1826%	
IBK (95 neighbors)	33.1008%	32.785%

Class Attribute: GENDER		
Algorithm	**Training Accuracy**	**Holdout Accuracy**
OneR	66.6171%	
J48 (conf. .29, m 19)	69.4034%	69.3659%
Naive Bayes	67.4172%	
IBK (61 neighbors)	70.7765%	70.6108%

Figure 4: Accuracy on Dataset 2

Dataset 3 Results

Class Attribute: LFSTAT		
Algorithm	**Training Accuracy**	**Holdout Accuracy**
OneR	84.4791%	84.6679%
J48	84.9025%	85.0233%
Naive Bayes	81.8265%	81.5664%
IBK (5 neighbors)	84.1329%	84.271%
IBK (10 neighbors)	84.1791%	84.2341%
IBK (20 neighbors)	84.1629%	84.1879%

Dataset 3 (Redundant Attributes Removed) Results

Class Attribute: LFSTAT		
Algorithm	**Training Accuracy**	**Holdout Accuracy**
Naive Bayes	83.3668%	83.5556%
IBK (5 neighbors)	84.6014%	84.631%
IBK (10 neighbors)	84.6106%	84.631%
IBK (20 neighbors)	84.5517%	84.5987%

Figure 5: Accuracy on Dataset 3

the original algorithms classified anything as 2, and Naive Bayes, and IBK (5 neighbors) classified 6 and 1 instances of class 2 correctly. This implies that unemployed respondents vary; they are not similar in aspects such as degree type, degree field, gender, or any of our other attributes. This explains why the accuracy of IBK generally decrease as our number of nearest neighbors increases. Because there are not very many instances in class 2, and because they are not concentrated, that means that the more neighbors

there are, the less likely it is that the majority of neighbors are class 2. We do see some improvement in accuracy after removing redundant attributes, but not much. This could be in part because our accuracy was fairly good to begin with, leaving less room for improvement. Or it could simply mean that there is no better way to predict a persons labor force status. When we look at association rules, we see that get trivial results about ethnicity and citizenship. However, it is still interesting to note that Weka has figured out these connections from data, rather than from being told. After removing the redundant attributes, we noted that in four out of the first five rules, the antecedent of the rule deals with LFSTAT, or labor force status. With no redundant attributes, there are fewer connections between the various attributes to make, so it is actually easiest for Weka to predict LFSTAT.

Dataset 4 Results: Figure 6 shows the accuracy of the various models run on Dataset 4, along with the correlation coefficient for linear regression (denoted by *). IBK with 20 neighbors gives the greatest accuracy, at 74.3%, closely followed by IBK with 10 neighbors. shows that people who rate satisfaction with specific aspects of their jobs in a similar manner also rate overall job satisfaction in a similar manner. This makes intuitive sense, as we expected peoples satisfaction levels to be consistent across all aspects. So, if two people had differed in opinion about one type of satisfaction, their total job satisfaction would have likely differed as well. Overall, respondents were satisfied with their jobs. Figure 7 shows the Confusion Matrix for the Naive Bayes algorithm. From it, we can see that 1, the highest ranking (4 is lowest) and 2 were the most common responses, making it most likely for these two to be misclassified as each other. As we move down the rankings, there are fewer and fewer misclassified instances. We get further evidence that most respondents are satisfied when we look at the association rules. For the first 100 rules, every single rules applies solely to category 1. Because it is the most common response, these rules have the most support. Their high accuracy shows the relationship between satisfaction in one area of a job and satisfaction in another. Lastly, we ran linear regression on this data. The correlation was .7573 for the following equation. We were hoping for the correlation coefficient to be closer to 1, but were unable to get it there. As every variable in the equation was positive, it did not look like we had redundant attributes that could be removed, and looking at smaller subsets of the attributes did not improve the linearity either. For each category, as well as overall satisfaction, we visualized the average response regarding satisfaction level. The column labeled 'Average Job Satisfaction' is the average response of all categories other than overall satisfaction. This visualization can be seen in Figure 8

Dataset 4 Results

Class Attribute: JOBSATIS		
Algorithm	**Training Accuracy**	**Holdout Accuracy**
OneR	67.4369%	67.4874%
J48	73.8117%	73.9232%
Naive Bayes	74.2549%	74.1955%
IBK (5 neighbors)	73.8047%	74.0066%
IBK (10 neighbors)	74.1048%	74.2844%
IBK (20 neighbors)	74.1368%	74.2956%
Numeric Estimation	.7597*	.7573*

Figure 6: Accuracy on Dataset 4

	Classified as 1	Classified as 2	Classified as 3	Classified as 4
Class 1	6595	1464	29	23
Class 2	1578	5797	589	64
Class 3	64	519	712	160
Class 4	13	41	99	246

Figure 7: Confusion Matrix for Naive Bayes Predicting Job Satisfaction

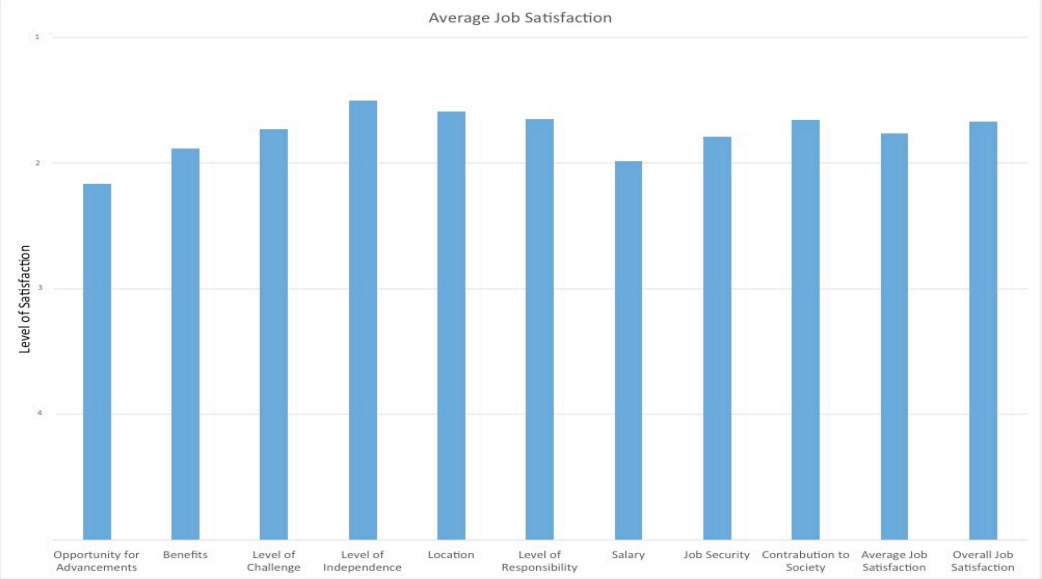

Figure 8: Average Job Satisfaction

6 Conclusion

Although we did not definitively determine who gets the job, we were able to better understand certain aspects of who has jobs and details regarding their job relevance and satisfaction. Looking further into the dataset we determined that those who responded as not being employed often did not answer any of the other questions. This may be because many of the questions regarded details of a presumed job. We were able to better understand the attributes that affect a person with a job. We produced predictive models for gender, salary, age, job satisfaction and labor force status. After learning that K Nearest Neighbors and J48 were the best models we determined that the data worked best when we did not assume that all attributes were not equally weighted. While this information is helpful in determining models predicting salary and job satisfaction, it does not conclusively advise in the best method of acquiring a job. In the future, we hope to find a dataset that has more detailed information about those who are unemployed.

Appendix

Attribute Name	Type*	Values	Description
ACTCAP	nominal	{N,Y}	primary or secondary work: computer applications
ACTDED	nominal	{N,Y}	primary or secondary work: development or design
ACTMGT	nominal	{N,Y}	primary or secondary work: management/sales
ACTRD	nominal	{N,Y}	primary or secondary work: basic research/applied research/development/design
ACTRDT	nominal	{N,Y}	primary or secondary work: basic research/applied research/development/design/teaching
ACTRES	nominal	{N,Y}	primary or secondary work: basic research/applied research
ACTTCH	nominal	{N,Y}	primary or secondary work: teaching
AGE	numeric	23-75	age
BACHELOR1	nominal	{1955,1960,1965, 1970,1975,1980, 1985,1990,1995, 2000,2005,2006}	year of first bachelor's degree
BTHUS	nominal	{N,Y}	place of Birth (US or non-US)
CH1218	nominal	{N,Y}	children living in household: ages 12-18
CHUN12	nominal	{N,Y}	children living in household: under 12 years
CTZUSIN	nominal	{N,Y}	citizenship status
HIGHDEGREE	nominal	{1,2,3,4}	highest degree type
GENDER	nominal	{F,M}	gender
GOVSUP	nominal	{N,Y}	whether an individual received government support
YRHIGHDEGREE	nominal	{1955,1960,1965, 1970,1975,1980, 1985,1990,1995, 2000,2005,2006}	year highest degree was earned
HRSWK	nominal	{1,2,3,4}	typical hours worked per week
JOBINS	nominal	{N,Y}	available benefits: health insurance
JOBPENS	nominal	{N,Y}	available benefits: pension/retirement plan
JOBPROFT	nominal	{N,Y}	available benefits: profit-sharing plan

ADVANCE	nominal	{1,2,3,4}	job satisfaction
BENEFITS	nominal	{1,2,3,4}	satisfaction with benefits job offers
CHALLENGE	nominal	{1,2,3,4}	satisfaction with challenge job offers
INDEPENDENCE	nominal	{1,2,3,4}	satisfaction with independence job offers
LOCATION	nominal	{1,2,3,4}	satisfaction with location of job
RESPONSIBILITY	nominal	{1,2,3,4}	satisfaction with responsibility job offers
SALARYSAT	nominal	{1,2,3,4}	satisfaction with salary
SECURITY	nominal	{1,2,3,4}	satisfaction with security of job
SOCIETY	nominal	{1,2,3,4}	satisfaction with job's contribution to society
JOBSATIS	nominal	{N,Y}	job satisfaction
JOBVAC	nominal	{N,Y}	available benefits: paid vacation/sick/personal days
MAJORSCIENCE	nominal	{N,Y}	job required technical expertise at bachelor's level or higher in: eng, comp sci, math, nat sciences
MAJOROTHER	nominal	{N,Y}	job required technical expertise at bachelor's level or higher in: other fields
MAJORSOC	nominal	{N,Y}	job required technical expertise at bachelor's level or higher in: social sciences
MINORITY	nominal	{N,Y}	minority indicator
MRDGYR	nominal	{1955,1960,1965, 1970,1975,1980, 1985,1990,1995, 2000,2005,2006}	year of most recent degree (5-year intervals)
MRDG	nominal	{1,2,3,4,5}	field of study for first bachelor's degree (major group)
JOBCODEGEN	nominal	{1,2,3,4,5,6,7}	job code for principal job
JOBCODESPEC	nominal	{318730.0,29889S, 39899S,43899S, 58799S,41929S, 22639S,42929S, 547280.0,33878S, 45939S,537260.0, 567350.0,527250.0, 44999S,19889S, 61199S,79999S, 71999S,69999S}	more specific job code
MAJOR1	nominal	{318730.0,29889S, 39899S,43899S, 58799S,41929S, 22639S,42929S,	major of first bachelor's degree

		547280.0,33878S, 45939S,537260.0, 567350.0,527250.0, 44999S,19889S, 61199S,79999S, 71999S,69999S}	
FIELD1	nominal	{1,2,3,4,5,6,7}	field of study for first bachelor's degree (major group)
HIGHMAJOR	nominal	{318730.0,29889S, 39899S,43899S, 58799S,41929S, 22639S,42929S, 547280.0,33878S, 45939S,537260.0, 567350.0,527250.0, 44999S,19889S, 61199S,79999S, 71999S,69999S}	field of major for highest degree
HIGHFIELD	nominal	{1,2,3,4,5,6,7}	field of study for highest degree
NEWBUS	nominal	{N,Y}	are you working for a business that came into being within the past 5 years
MRMAJOR	nominal	{318730.0,29889S, 39899S,43899S, 58799S,41929S, 22639S,42929S, 547280.0,33878S, 45939S,537260.0, 567350.0,527250.0, 44999S,19889S, 61199S,79999S, 71999S,69999S}	field of major for most recent degree
MRFIELD	nominal	{1,2,3,4,5,6,7,9}	field of study for most recent degree
RELATEDDEGREE	nominal	{1,2,3}	extent that principal job is related to highest degree
MEMBERSHIPS	nominal	{0,1,2,3,4,5,6}	number of professional society memberships
INVOLVEMENT	nominal	{N,Y}	attended a society meeting in the last year
RACETH	nominal	{1,5,7}	race/ethnicity
SALARY	numeric	0-150000	salary
SUPERVISOR	nominal	{N,Y}	supervised others during reference week
WAPRI	nominal	{1,2,3,4,5,6,7,8,9, 10,11,12,13,14}	work activity spent most hours on in principal job
WASEC	nominal	{0,1,2,3,4,5,6,7,8,9, 10,11,12,13,14}	work activity spent second most hours on in principal job
WKSWORKED	numeric	1-52	weeks per year worked in principal job

TRAINING	nominal	{N,Y}	attended work-related training
LFSTAT	nominal	{1,2,3}	labor force status (1 = employed, 2 = unemployed, 3 = not in labor force)

*the attribute's original type; numeric values were often discretized during analysis
**information regarding what the values of nominal attributes represent is included in the file with the original data referenced in the Dataset Description section