William Pembleton

John Doucette

Data Mining & Machine Learning

11/28/2018

What causes Job/Career satisfaction in Computer Science and would you write unethical software?

Abstract

This report takes a look into what variables cause job satisfaction, career satisfaction, and if a user would write unethical software. Surprisingly, good habits like keeping a good sleep schedule, not skipping meals, and having an exercise routine didn't show up in any of the trees. However, I think that this is due to me leaving in career satisfaction when I was building the job satisfaction tree and vice versa. This is a place where future research can be done.

Background

The data I decided to use for this project is from the website Stack Overflow. Stack Overflow is an online community for developers to learn, share their knowledge, and develop their careers. It's essentially a Q&A website for developers similar to something like Yahoo Answers. Stack Overflow also does a yearly survey asking their users to answer questions about their job. They asked questions like "How satisfied are you with your current job?", "On a typical day, how much time do you spend outside?" and "What is your current gross salary (before taxes and deductions)?".

What I was interested in studying was which variables lead to a high job satisfaction and a few other things. What that means is this question is a supervised learning problem. In a supervised learning problem, you have all the data and you have the most important attribute (the class), but you don't know what is causing the class. I decided to solve this problem with the tree-based algorithm J48.

J48 is a particular implementation of the Decision Tree Learner. The Decision Tree Learner is a generic algorithm that creates trees and leaves how to pick each attribute to the concrete implementations. J48 chooses attributes based on information gain. Information gain is calculated by taking the entropy of a set and subtracting the sum of the entropy of the subsets split on a feature.

ID3 was the original implementation of a decision tree and it was created by John Ross Quinlan in 1986. There were iterative developments from ID3 to C4.5 and C4.5 to C4.8 from there the developers of WEKA (The software suite I'm using to run these tests) modified the C4.8 algorithm to work in Java thus J48.

The problem with making decision trees is that the only way you can test how accurate the tree is is to run the examples through the tree. The main problem is with that once you run testing data through your tree and use that data to make changes in your model then you've basically spoiled that testing data and you need to go collect more. Instead, if you were to make k

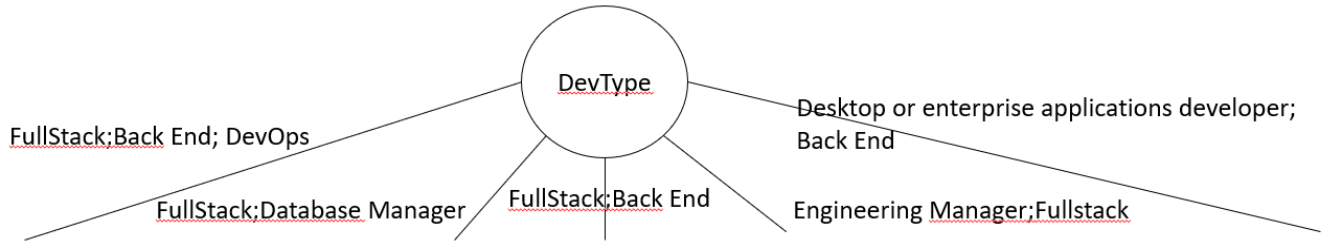
partitions in your dataset and trained on every partition to build a model except for one then you could use that last partition to test your model. Then you could pick another partition that you haven't chosen yet and train on the remaining partitions. Repeating this for k partitions gives you k models and an accuracy for each model. Then you could average the trees and the accuracy of each model and thus not spoil any data. This is known as K-Fold Cross-Validation. For my project, I used 10-Fold Cross-Validation.

A common problem that comes up with decision trees is that they will keep splitting until they run into pure subsets or it runs out of attributes to split on. When that happens, the tree may grow very deep and only gain a trivial amount of accuracy, but not only that, it would reduce the accuracy on data the model has not seen. This problem is solved via pruning the tree which means to cut it off once some condition is true. For historical reasons the way to do this is with a confidence factor which is basically a p-value. So for each node, the algorithm does a statistical test to determine if it should prune this node. If this node fails the test then that node is pruned and a leaf is put in it's place. The leaf's value is the value of the class that shows up most often in this subset of the dataset.

Methodology

There was a large number of attributes that needed to be discarded because they didn't help answer the question at hand.

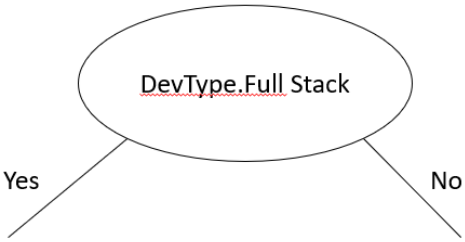
In the preprocessing phase, I needed to create binary features for some attributes in the data. For instance, on the question "What type of developer are you? Select all that apply" the data came in as "Full Stack;Back End;DevOps". This would cause a problem with a normal implementation of J48 because each row would be treated as their own path for a node to go down.



This would cause some obvious problems.

- If one wanted to interpret the tree one would have to find their exact qualifications in the tree and follow that branch (one might have to search through hundreds of unique combinations to find their branch).
- The tree would grow very wide which would significantly slow down training time.

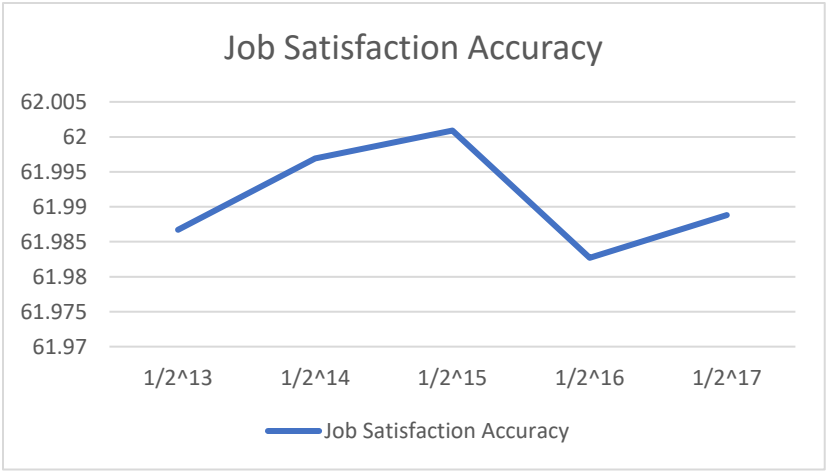
I solved this problem by processing every row that had semicolons in them into binary features. As an example, for DevType for each value, I made a column that had either Yes or No if a particular example was that type of developer.



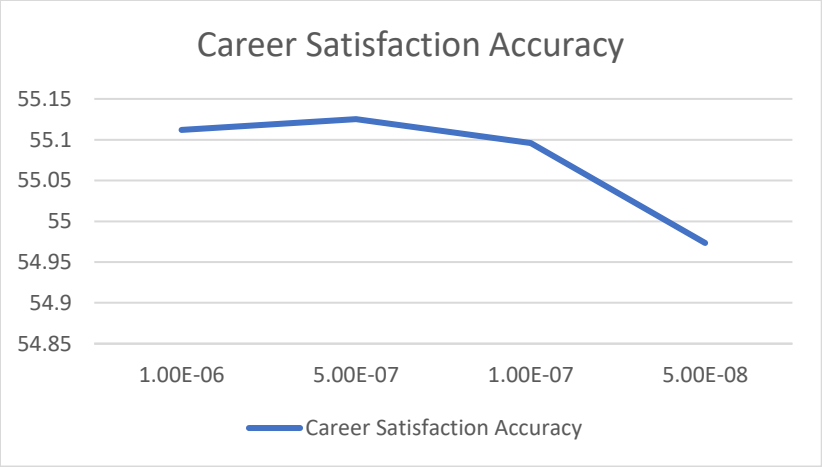
There were a few other small changes that had to be made to the data so that the model would work in WEKA.

- The attribute ConvertedSalary (which was the developer's salary converted to USD at the exchange rate on 1/18/2018) needed to be binned. This was done so that the tree doesn't grow wide very quickly, and the results are more applicable to a broader audience.
- I had an attribute NumberMonitors which is the number of monitors a user had at their workstation. I needed to convert the numerical values (1, 2, 3) to words (One, Two, Three) so that WEKA would see this attribute as a nominal value (categorical) instead of a numerical one.
- I deleted the attribute Country because I wanted to make this tree globally applicable. There was another attribute Currency which was the type of money a person is paid out in (USD, British pound, Euro) for the same reason that it would tie an example to a country.

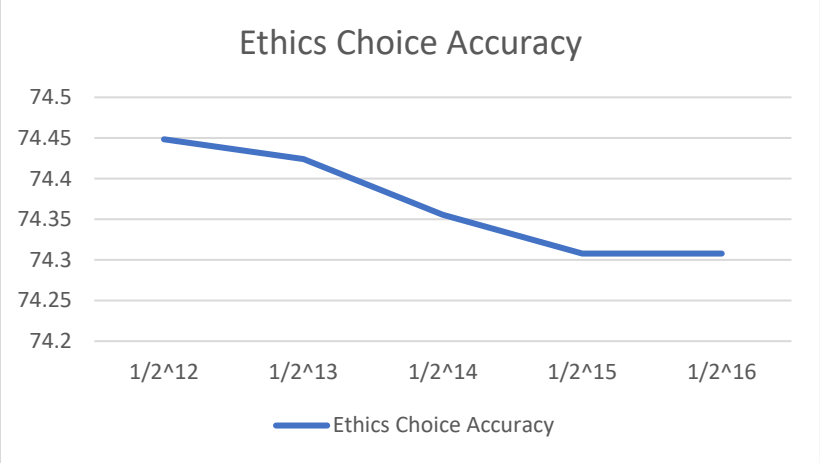
For the job satisfaction tree, I chose a Confidence Factor of 1/2^15 because I found it to have the highest accuracy of all the Confidence Factors I tried.



For career satisfaction, I chose a confidence factor of 1E-7 because the accuracy isn't reduced by much and it's the first tree that isn't finding patterns in random noise.



For ethical choice, I chose a Confidence Factor of 1/2^14 for the same reason of it finding patterns in random noise.



Results

Because the trees for this data are way too wide to be seen on this report I decided to make a link to where you can download the trees and look at them for yourself. I also give a short description of what each tree looks like in case you don't want to go look at them for yourself. When you're reading through the trees the number on the left is the number of examples that went down this path and the number on the right is the number of examples the tree guessed incorrectly on this path.

https://tinyurl.com/yalmfzlg

First off, the biggest factor in predicting how satisfied you are with your job is if you use some kind of communication software.

> Once the tree is split on that the next row is how satisfied you are with your career. If you went down the path of using a communication tool then the row is JobSearchStatus (Which of the following best describes your current job-seeking status?).

From here the tree splits on the type of employment you have (Full time, part time etc.) from there it goes to leaves.

If you went down the path of not using a communication tool the tree splits on how satisfied you are with your career, from here it goes to small branches but mostly leaves.

What's more interesting is what doesn't show up in the tree rather than what did show up.

- For instance, Age didn't show up at all which tells me that how satisfied you are with your job has little to do with how old you are.
- Open source shows up but only rarely which tells me that if you work on open source projects it doesn't affect how satisfied you are with your job.
- How long you've been coding both professionally and non-professional doesn't show up.
- wakeTime, hoursComputer, exercise, skipMeals (In a typical week, how many times do you skip a meal in order to be more productive?) didn't show up at all which I found very interesting because I expected that doing healthy habits would affect how satisfied you are with your job.

```
=== Confusion Matrix ===

    a    b     c    d    e     f    g    h   <-- classified as
 4383  146  7566  112   20     1  108  100 |     a = Extremely satisfied
   58 2212  1855  442  104     3 1060  584 |     b = Moderately dissatisfied
 2095  547 21402  788  144     2  861  166 |     c = Moderately satisfied
   80  394  2270 1232  112     0  791   87 |     d = Neither satisfied nor dissatisfied
  211  457  7108  878  246     1 1012   99 |     e = Slightly satisfied
   73   50   273   60   16 29028   42   37 |     f = NA
  103  935  3240  718  153     0 1729  179 |     g = Slightly dissatisfied
   60  602   537   90   65     4  245  879 |     h = Extremely dissatisfied
```

Like the job satisfaction tree, the first attribute the Career satisfaction tree chose to split on is whether you use a communication tool. Something to note is that when people answered the question about what communication tool they use they were probably thinking about the current job that they're in rather than all communication tools they used in their career. The question is phrased as "Which of the following tools do you use to communicate, coordinate, or share knowledge with your co-workers? Please select all that apply."

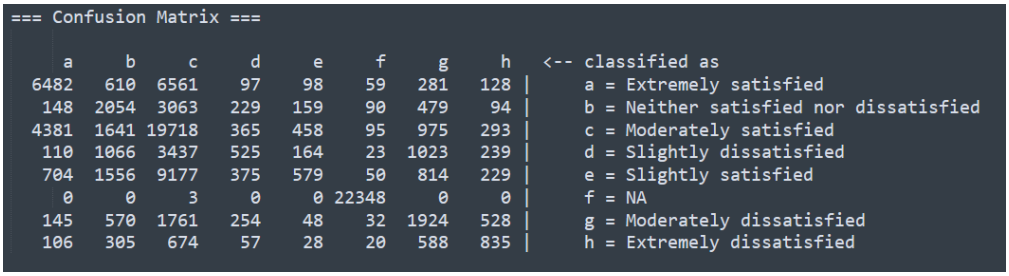
From there if you're using a communication tool then the next factor the tree split on is how satisfied you are with your job. If you have high job satisfaction, then you likely have high career satisfaction. There are a few other branches but they're for minor leaves.

If you're not using a communication tool then the next biggest factor is what you hope to be doing in the next 5 years (I think J48 chose this because it will split the tree into a bunch of sub trees not because there's some kind of underlying relationship).

After those splits, 5/8 of them split on job satisfaction and the other 3 split on some other variable because the number of examples they hold is usually < 500 while the ones the split on job satisfaction vary in their number of examples from 1,000-24,000.

Once again, its the stuff that didn't show up in the tree that's the most interesting.

- Age didn't show up again which is good to hear because how old you are seemingly doesn't affect how satisfied you are with your career.
- Language shows up this time but only in circumstances where the number of examples is small (usually < 300).
- Open source doesn't show up in this tree which I expected to show up more in career satisfaction than job satisfaction but for some reason it didn't.
- How long you've been coding both professionally and non-professionally only showed up once.
- wakeTime, hoursComputer, exercise, skipMeals didn't show up again which I found very interesting because I expected that doing healthy habits would affect how satisfied you are with your career more so than your job.



```
=== Confusion Matrix ===

     a     b     c     d     e     f     g     h   <-- classified as
  6482   610  6561    97    98    59   281   128 |     a = Extremely satisfied
   148  2054  3063   229   159    90   479    94 |     b = Neither satisfied nor dissatisfied
  4381  1641 19718   365   458    95   975   293 |     c = Moderately satisfied
   110  1066  3437   525   164    23  1023   239 |     d = Slightly dissatisfied
   704  1556  9177   375   579    50   814   229 |     e = Slightly satisfied
     0     0     3     0     0 22348     0     0 |     f = NA
   145   570  1761   254    48    32  1924   528 |     g = Moderately dissatisfied
   106   305   674    57    28    20   588   835 |     h = Extremely dissatisfied
```

I would suggest going to look at the tree this time, the tree is pretty small.

EthicsChoice: Imagine that you were asked to write code for a purpose or product that you consider extremely unethical. Do you write the code anyway?

EthicsReport: Do you report or otherwise call out the unethical code in question?

EthicalImplications: Do you believe that you have an obligation to consider the ethical implications of the code that you write?

I was rather surprised by the number of people that said that they feel an obligation to consider the ethical implications of the code that they write but they didn't say that they wouldn't write the software. A look at the confusion matrix though shows that most times that the tree classified a "No" and there was a mistake the most common answer was "Depends on what it is". Which makes sense because it means that people will consider the software that they write and decide to write the software if they feel if it is ethical or not. Nearly 2,000 people went down the path saying that you don't have an obligation to consider the ethical implications of the code you write but think that they will report the code depending on what it is. I think the main problem for people of this group is that they feel the word obligation is too strong and they will report depending on how severe the ethical implications of the software is.

```
=== Confusion Matrix ===

     a     b     c     d   <-- classified as
 36932  4184   108   217 |     a = No
 17568  8114    68   181 |     b = Depends on what it is
    67    21 27985     0 |     c = NA
  1936   984    17   473 |     d = Yes
```

A rather large caveat to this research is that all the data that was collected was self-reported which leads to a few downsides. One of them is that some people will just click through the survey when they get bored selecting random answers as they go or some will leave without completing the entire survey. Another one is that some people might lie on the survey or choose an answer they wouldn't actually do in real life.

Future research

There are a few different areas that future research into this topic can go.

1. Look at the times where people are satisfied with their career and not satisfied with their job and vice versa and see if there is something that causes a person to be happy with their career but not their job.
2. Building a tree without career satisfaction for the job satisfaction tree and seeing how much accuracy changes. Likewise, do so for the career satisfaction tree. The reason for doing this is to see if variables that have to do with good habits (exercise, not skipping meals etc.) show up and how important they are they to predicting job/career satisfaction.
3. Do a full study and ask participants what they think causes their job/career satisfaction.