William Pembleton

11/12/18

What causes job/career satisfaction in Computer Science and would you write unethical software?

Abstract

To be filled in later approximately 200 words talking about the methodology and results of the paper. I previously was making my own implementation for ID3 and that turned out to be unfruitful. So now I’m going to be using scikit-learn’s implementation of ID3. This section will be filled in once the processing has been completed.

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 in their careers. It’s essentially a Q&A website for developers similar to something like Yahoo Answers. Stack Overflow also does a yearly surveys 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, but you don’t know what is causing that attribute. I decided to solve this problem with the tree-based algorithm J48.

J48 is based on C4.8 which was based on C4.5 which was based on ID3. ID3 was the original implementation of a decision tree and it was created by John Ross Quinlan in 1986. There was 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.

J48 is a particular implementation of a Decision Tree Learner. The Decision Tree Learner is a generic algorithm that creates trees and leaves it open to the particular implementation how to choose each attribute. 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.

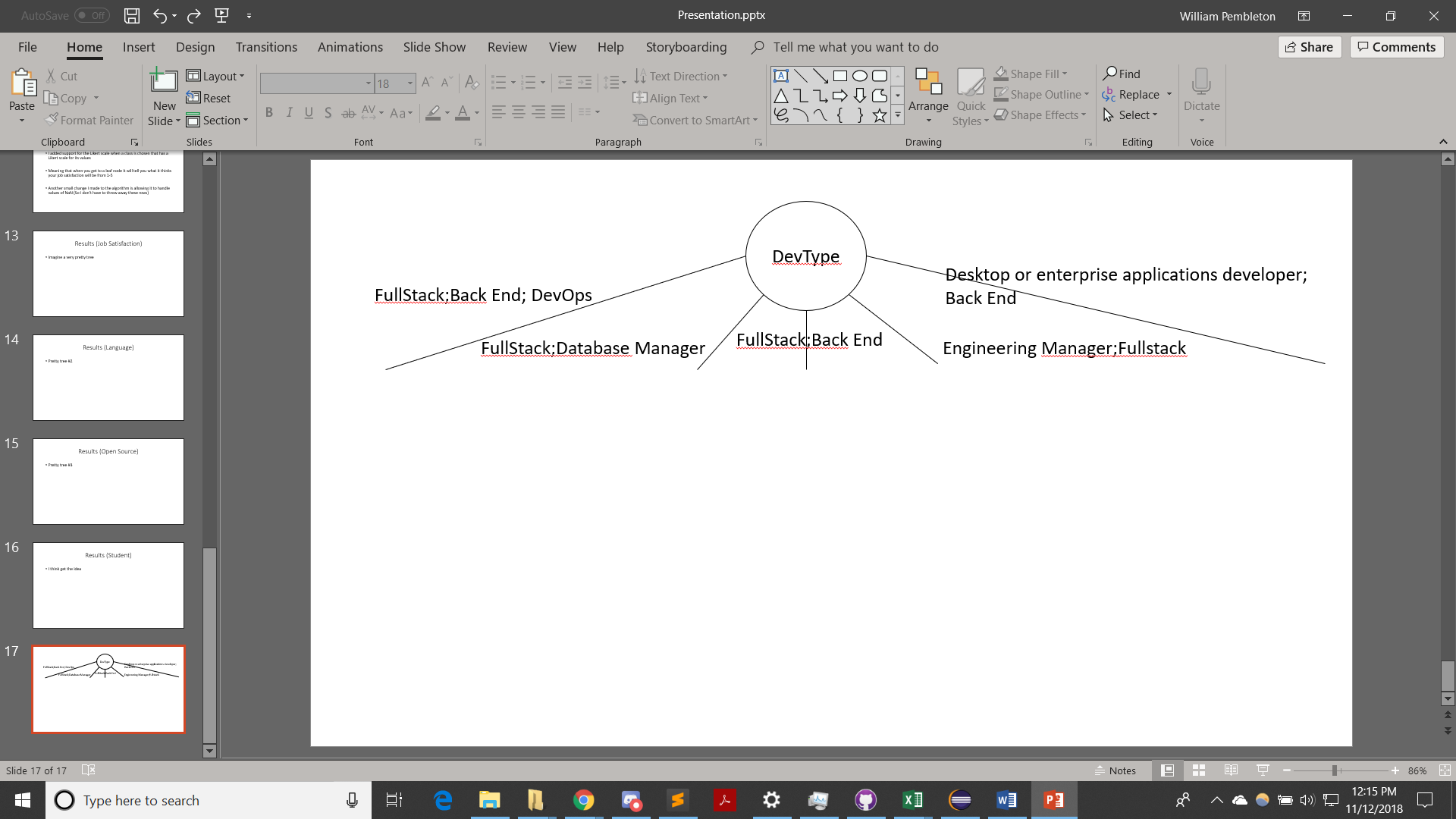
**K-Fold cross validation**

**Confidence Factor**

Methodology

I got rid of a lot of attributes either because they weren’t useful to answering the question I proposed or because the attributes were asking things to help Stack Overflow.

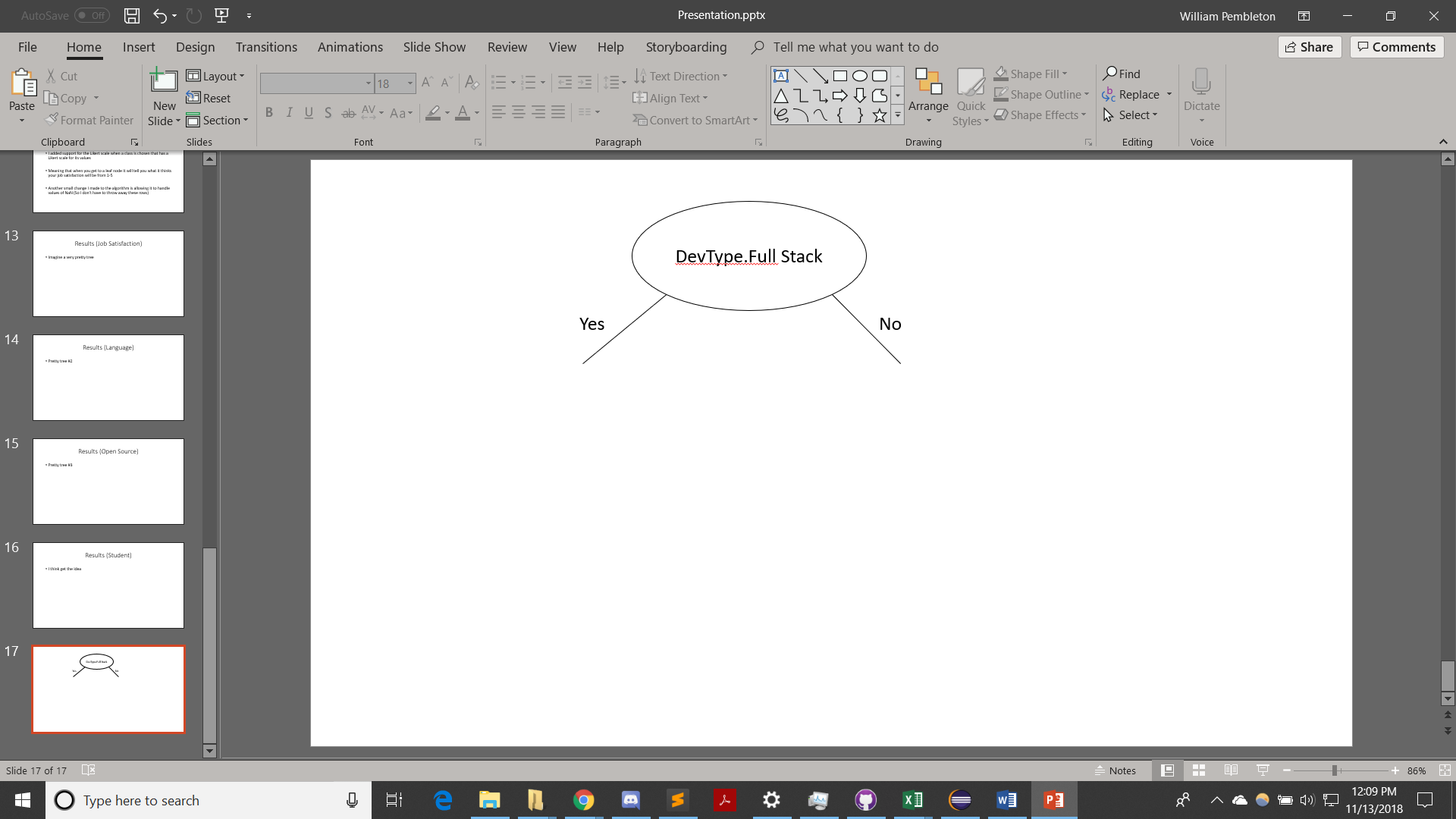
In the preprocessing phase I needed to parse the data. For instance, on the question “What type of developer are you? Select all that apply” the data came in as “Full Stack;Back End; DevOps”. This would cause a problem with a normal implementation of ID3 because each row would be treated as their own path for a node to go down.



This would cause some obvious problems.

* If you wanted to use the tree you would have to find your exact qualifications and follow that branch (which you might have search through the hundreds of unique combinations to find your branch)
* The tree would grow very wide which would significantly slow down training time

I solved this problem by preprocessing every row that had semicolons in them into their own columns. So there would be a column for every unique value within that row. As an example, I would make a row for Full Stack if we’re talking about what type of developer one is and in this row the values are either Yes or No which if a particular row contains Full Stack then this new column has a Yes in it.



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

* I binned the attribute ConvertedSalary (which was the developer’s salary converted to USD at the exchange rate on 1/18/2018). 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 their work station. I needed to convert the numerical values (1,2,3) to word values (One, Two Three) so that Weka say 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 payed out in (USD, British pound, Euro) for the same reason that it would tie a example to a country.

For the job satisfaction tree I used a Confidence Factor of 1/(2^15) because I found it to have a slightly higher accuracy than all other Confidence Factors I tried.

For Career satisfaction I chose a confidence interval of 1/(2^10) for the same reason.

For Ethical Choice I chose Confidence Factor of

Results

**https://tinyurl.com/yalmfzlg**

Because the trees for this data are way too wide to be seen on this report I decided to make a link to Google Drive 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.

First off the biggest factor to predicting how satisfied you are with your job is if you use a communication software. Once the tree is split on that the next row is how satisfied 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?) which makes sense, if one is not satisfied with their job then they’re likely looking for a new one and thus have a low JobSatisfaction score or if you’re happy with your job then you’ll likely not be open to a new job opportunity. If you went down the path of not using a communication tool the tree split on the type of employment you have (Full time, part time etc.) then goes to leaves.

What’s interesting is that quite a few variables didn’t show up much/ at all.

* For instance, Age didn’t show at all which tells me that how satisfied with your job has little to do with how old you are
* Language doesn’t show up
* Open source shows up but only rarely which tells me that if you commit to open source 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.

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 coworkers? 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 it splits on what your hope is in the next five years. I think this was chosen because it makes a lot of sub trees and not because there’s some kind of 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 are usually < 500 while the ones the split on job satisfaction vary in their number of examples from 1000-24,000

Once again it makes sense that career satisfaction and job satisfaction are related but once again what’s more interesting is what’s not listed in the tree

* Once again Age didn’t show up which is good to hear because those variables seem to not have any connection
* 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 so in career satisfaction than job satisfaction but for some reason it didn’t
* How long you’ve been coding both professionally and non-professional only showed up once
* 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.

<https://s3.amazonaws.com/academia.edu.documents/30661154/HerzbergJobSatisfaction_JMP_2009.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1543331702&Signature=4VFm4iBlToe1Ndg6xm9CsXQCSt8%3D&response-content-disposition=inline%3B%20filename%3DPersonality_motivation_and_job_satisfact.pdf>

Results demonstrate that between 9 and 15 per cent of the variance in motivation is accounted for by demographic variables and the Big Five personality traits. In line with previous findings (Judge et al.), conscientiousness and job status were both significant predictors of job satisfaction, and between 11 and 13 per cent of the variance was accounted for by personality and other demographic variables

<https://www.researchgate.net/profile/Debra_Major/publication/228358052_Comparing_Telework_Locations_and_Traditional_Work_Arrangements_Differences_in_Work-Life_Balance_Support_Job_Satisfaction_and_Inclusion/links/53e501b50cf25d674e953076/Comparing-Telework-Locations-and-Traditional-Work-Arrangements-Differences-in-Work-Life-Balance-Support-Job-Satisfaction-and-Inclusion.pdf>

Main office workers reported higher levels of WLB support than satellite and client-based workers. Additionally, main office workers reported the highest levels of workplace inclusion.

<https://bmchealthservres.biomedcentral.com/articles/10.1186/1472-6963-14-466>

Job satisfaction doesn’t affect your work