William Pembleton

11/12/18

What causes job satisfaction in Computer Science?

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 build their careers. It's essentially the Q&A website for developers similar to something like Yahoo Answers. Stack Overflow also does 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)".

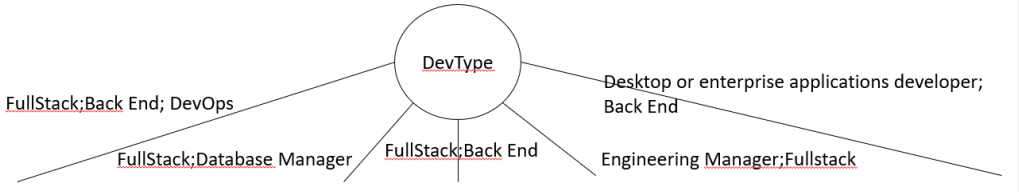
What I was interested in studying was which variables lead to a high job satisfaction. Which means that this question is a supervised learning problem. In a supervised learning problem you have all the data and you have the most important attribute, but you don't know what is causing that attribute. I decided to solve this problem with the tree-based algorithm ID3.

ID3 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. ID3 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.

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.



Commented [WP1]: Format the paper in some style

Commented [WP2]: I decided to go the hand wavy route with explaining how entropy is calculated. Let me know what you think

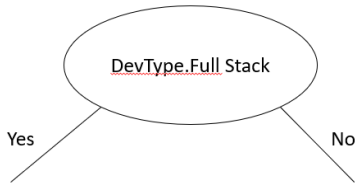
Commented [WP3]: Wording

Commented [WP4]: Phrase this better

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 small changes that had to be made

- For each value of Job Satisfaction I gave a numerical value so that the algorithm can easily calculate how satisfied with your job you are.
- 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.

Results

I plan on having trees for the following classes Job satisfaction, Career satisfaction and an answer to the question "Imagine that you were asked to write code for a purpose or product that you consider extremely unethical. Do you write the code anyway?"

I expect this section to have a picture of each tree on each page and discussion of each tree at the end of the paper.

Commented [WP5]: I don't think bullet points should belong in this paper, so change later

Commented [WP6]: Explain this better and more concisely

Commented [WP7]: Come up with a concise way to say this