William Pembleton

Dataset:

https://www.kaggle.com/stackoverflow/stack-overflow-2018-developer-survey#survey_results_public.csv

Abstract

To be filled in later

Background

The data I decided to use for this project is from the website Stack Overflow. Stack Overflow is a 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 what variables lead to a high job satisfaction. Which means that this question is a supervised learning problem. A supervised learning problem is you have all the data and you have the most important thing, but you don't know what is causing that most important thing. What I decided to do was to use the tree-based algorithm ID3.

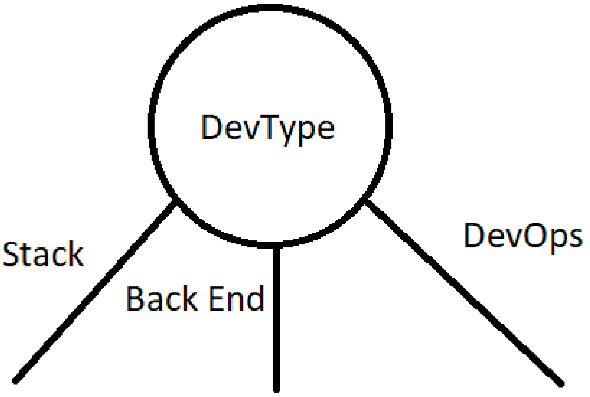
ID3 is a particular implementation of a Decision Tree Learner. The Decision Tree Learner is a generic algorithm that creates trees and leaves it open to the particular implementation how to choose each attribute. ID3 chooses attributes based on information gain. Information gain is calculated by taking the entropy of a set and subtracting the sum of the entropy of the subsets split on a feature.

Methodological

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". I separated each of these out into their own columns otherwise ID3 would treat each individual string within a feature as a path to branch down. Which would lead to a very wide and essentially useless tree.

DevType.Full Stack	DevType.Back End	DevType.Devops
Yes	Yes	Yes
Yes	No	Yes

There is some changes that are necessary to support this within the algorithm. I'm not exactly sure how I'm going to do it but I'm confident I'll find a way. In the end I want the tree to split on each of the values in the row



Once that's all completed I'll run the algorithm on the data

Graphs

I plan on having trees for the following classes Job satisfaction, What language a person uses, and if a person contributes to open source.

Conclusion