William Pembleton

John Doucette

Data Mining & Machine Learning

October 6, 2018

**Stack Overflow Developer Survey 2018**

Stack Overflow did a survey of people who use their site. In the survey they asked a little over 100 questions. I’m interested in seeing if and what the correlations are between the questions.

For instance, they asked questions like “How satisfied are you with your current job?” and “On a typical day, how much time do you spend outside?”. I’m interested in seeing what the correlations between answers are, but I’m really interested in is seeing how Job Satisfaction relates to variables like language, IDE, years spent professional coding, how often they skip meals and a few others.

I chose this dataset over something more fun because it would provide me information about the software industry I don’t know about yet. I came down to choosing between two datasets. One on trending YouTube videos from the past months and the Stack Overflow dataset. My questions for the YouTube dataset would have been “What tags are common for trending videos?” or “Can I predict what YouTube videos will be popular?”. The second question is more interesting though because I’m going to use a supervised learning model. However, I decided to not use the YouTube dataset because I’m not going to learn something of use to me from it.

As always, some data is missing so I’m going to put in some type of data that my implementation of the algorithm will ignore. I may need to use fewer categories than what was originally applied in the data. So that the rules have a chance to pass through the filter. However, I suspect I won’t need to do this because there was nearly 100,000 people who filled out the survey. Other than that there isn’t going to be much preprocessing that needs to be done. I’m thinking I can use Apriori to answer my question. Though there might be more interesting algorithms that can answer my question that I don’t know about.