It was briefly mentioned during this class but there was moment we talked about confounding. From my understanding confounding is an issue when features contributing to some output variable are interrelated to a point that it causes issues in the model. The reason for this issue is that if two variables are very heavily correlated and they both exist in the model you’re essentially doubly-accounting for a single root cause. I’m having this very issue at work – I’m performing natural language processing on PDFs to extract features from credit reports. Say I have a credit agreement that is 200 pages long and all I want to know is the start date, end date, the fee structure, and applicable rates of the credit agreement. My tool parses the PDF and extracts the paragraphs that pertain to each one of these features. My issue is that I’m applying a bag-of-words methodology and it’s causing my dataset to be very wide. With this discussion of confounding I analyzed my input variables and found that many of them are perfectly correlated. For example – the feature “Hong” and the feature “Kong” are perfectly correlated because if a paragraph contains the word “Hong” then it almost certainly is followed by “Kong”. With this in mind should I featurize it such that I only have one feature for the phrase “Hong Kong”?

Thanks,

-Oliver