Walter Blair
DATA 101 Final

In our initial exploration, we were trying to find particular player attributes that were tightly correlated to match outcomes. We summed all players on a team to come up with overall team ratings for these attributes for each match.

The first thing we did was plug this handful of overall team ratings into a K-nearest neighbor classifier. We didn’t think too much about what type of cross-validation we used, so we just went with a 90%/10% holdout scheme without repeated subsampling. Toward the end of the semester, I tried out a 10-fold cross validation and got much worse accuracy. It’s bugging me that I don’t understand why, so this is my attempt to figure it out. I should note that I don’t fully understand k-fold stratification, and my folds weren’t stratified.

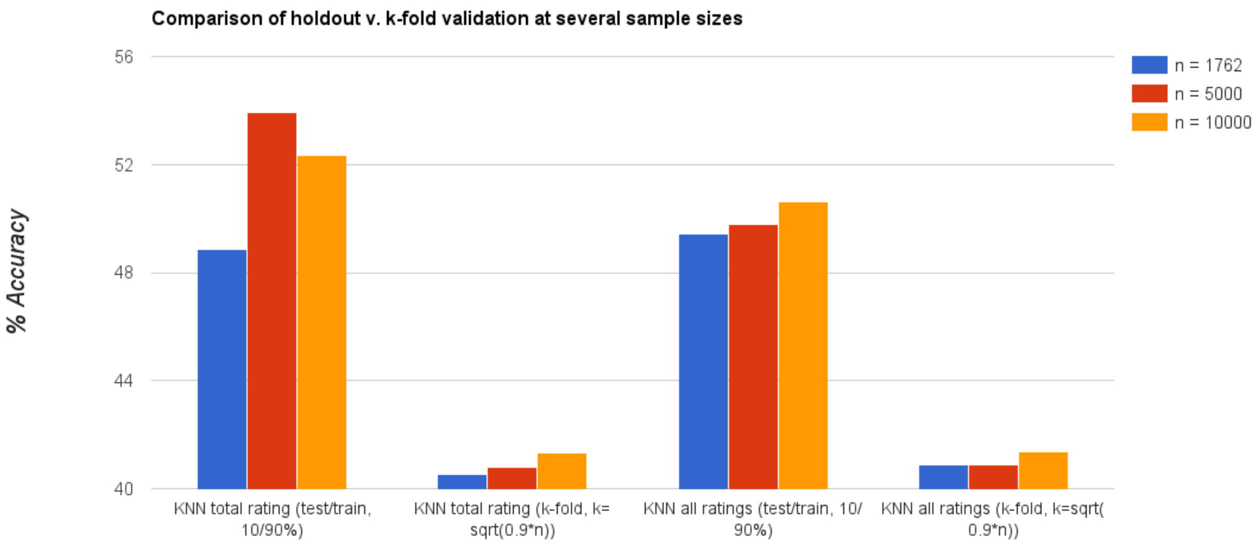

Figure 1. Note that sample size is varied by taking first n rows in dataframe, not by random sampling.

Rob Kohavi (1995) compared holdout and k-fold cross validation schemes, recommended stratification which I didn’t do, and found that k-fold validation reduces variance compared to holdout. The author also notes that repeated subsampling is nice, but my computer can only take so much.

I don’t yet totally understand the variance-bias dilemma, but it sounds like bias is sort of like a Type II error which is obtaining a false negative. It seems like bias will miss some important variation some of the time. Variance on the other hand sounds like Type I error, obtaining a false positive, because it will tend to find patterns that aren’t there.

So putting that together, if it’s true that k-fold validation tends to reduce variance and slightly increase bias relative to holdout, then that means that k-fold validation will underestimate the importance of some of the variation in the model. This might perfectly explain the surprising pattern here where k-fold validation results in a less accurate model. Our initial approach was to carefully select variables that we were pretty sure were important. In other words we were carefully pruning the big mess of data in order to put something that was almost certainly meaningful into the model. If we did a good job, then we wouldn’t want our model to be overly cautious in seeing patterns or it might miss all the good stuff in there. Perhaps this is what happened with k-fold validation – we picked a validation scheme that is overly cautious and underfit, trying to protect us from finding meaningless patterns in our data, but we had already tried to be cautious in cleaning the data so it ignored variation that would have been helpful in more accurately predicting match outcomes.

That’s my story and I’m sticking to it. Thanks for a fun semester!

References

Ron Kohavi. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. International Joint Conference on Artificial Intelligence (IJCAI). https://pdfs.semanticscholar.org/0be0/d781305750b37acb35fa187febd8db67bfcc.pdf

https://en.wikipedia.org/wiki/Bias%E2%80%93variance_tradeoff