Building Models with Jeff Sackmann’s match datasets

Jeff Sackmann’s ATP and WTA match databases contain player statistics and match summary statistics going back decades. We can build two types of models with this data:

# Explanatory Model

If we want to predict a match winner before the players step onto the court, we can only use information that is known prior to the match starting. This means, of course, that the match summary statistics are useless to us as predictive attributes, since before the match begins we would have no data to feed into the model.

We can, however, use the summary statistics to build an explanatory model. A successful explanatory model would tease out performance factors besides the players themselves that are reliably correlated with match outcomes.\* While not a predictive model, this model would be insightful as a learning tool on the sport itself.

\*The same performance factors could correlate differently with match outcomes depending on various match circumstances, such as court surface, racquet technology, and trends in the sport over time, so a prudent modeler would check for effects of these factors.

# Predictive Model