CS504 Final Project - Potential Code Changes

jsutor2 <jsutor2@masonlive.gmu.edu>

Thu 4/30/2020 9:58 PM

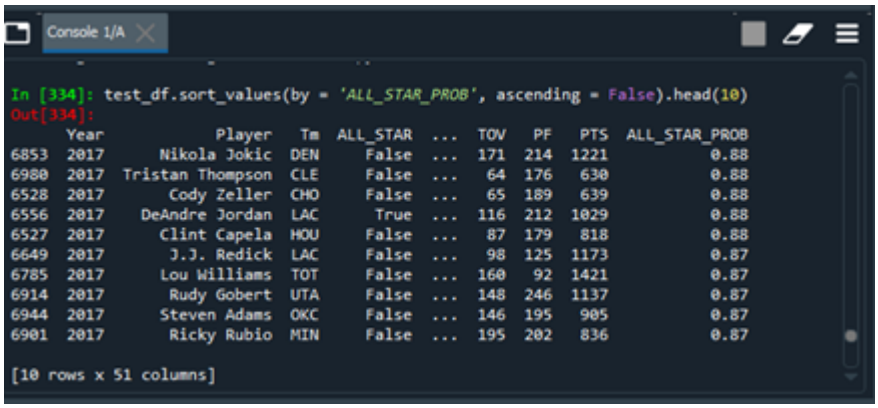
To: cricha26 <cricha26@masonlive.gmu.edu>; Pamir Rahimzadeh <prahimz2@gmu.edu>; wwashing <wwashing@masonlive.gmu.edu>

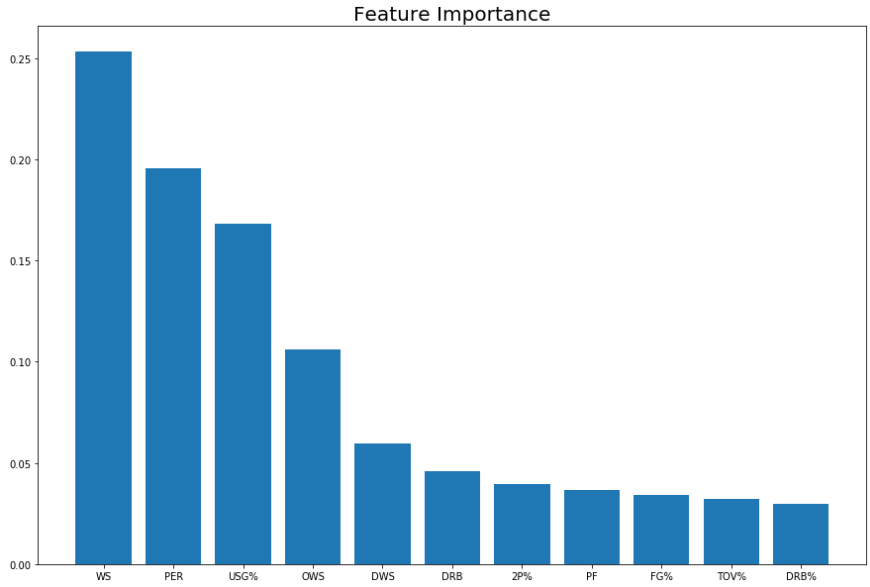
Hey guys,

This is kind of a long email, but essentially as I was testing our data again and trying to figure out the issue with everyone having a 100% All-Star prob I found some issues with the model. I think we need to limit our variables differently first and I also don't think the scaling was working correctly. I can upload the code to a new branch, but here is an explanation of what I found and the changes I think we need to make:

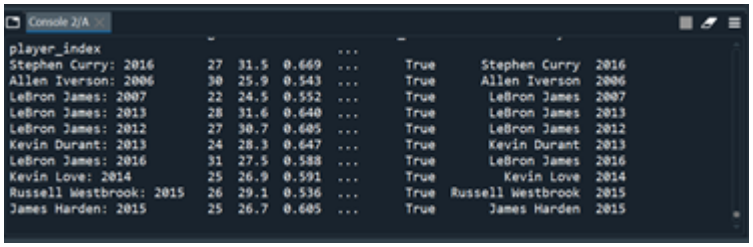
The scaling was making everyone have a 100% probability of being an All-Star in the training data, which meant that the test data was also messed up. See in the below image how only one of the players was actually an All-Star in 2017 but our model gave them the highest probability? I also noticed this because 7 of those players are centers because our model has both DRB and DRB% as two of the key features. This seems weird and since there is collinearity between these I feel like we need to limit our variables differently.

So I think we should only use % variables where those are available, and we should drop anything that is not completely atomic - meaning, we should include DRB% and ORB% but not include TRB% since that is just a sum of the other two. This way we'll eliminate both multicollinearity and non-atomicity. The one other change I made was to create a PPG variable of points divided by games (and dropped PTS from the dataset) - I think using this variable and the % variables will avoid the need for the scaling, although I am open to using some sort of scaling if we can figure out how to get it to work properly.

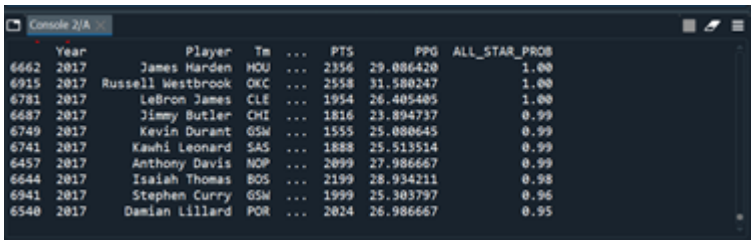




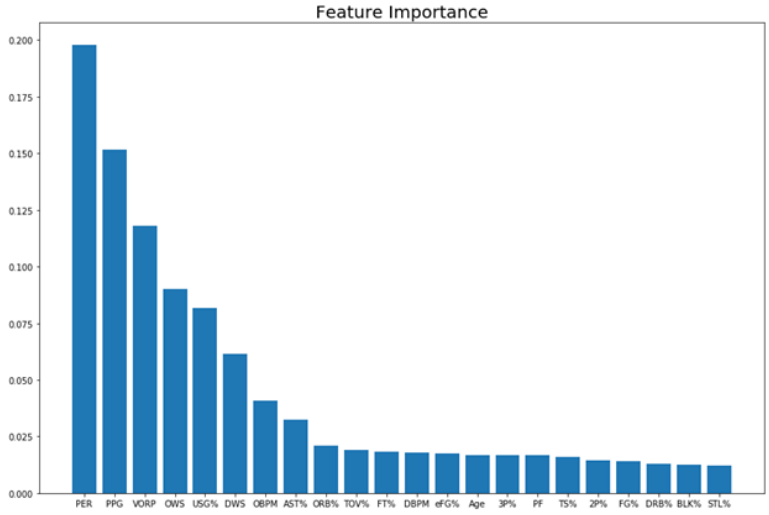
Using all of the remaining 22 variables, the model actually recommends retaining all of the variables and gives results that make more sense and performs better on the testing data. PPG is one of the top predictors, which makes more sense than it not even being part of our model. Also, in the training data, it gives Steph from 2016 the highest probability of being an All-Star which makes a lot of sense since that was when he was the first unanimous MVP. Then, in the 2017 training data, all ten players with the highest probability were in fact all stars - and actually Westbrook and Harden finished 1st and 2nd in MVP voting that year.



player_index					...			
Stephen Curry: 2016	27	31.5	0.669	...	True	Stephen Curry	2016
Allen Iverson: 2006	30	25.9	0.543	...	True	Allen Iverson	2006
LeBron James: 2007	22	24.5	0.552	...	True	LeBron James	2007
LeBron James: 2013	28	31.6	0.640	...	True	LeBron James	2013
LeBron James: 2012	27	30.7	0.605	...	True	LeBron James	2012
Kevin Durant: 2013	24	28.3	0.647	...	True	Kevin Durant	2013
LeBron James: 2016	31	27.5	0.588	...	True	LeBron James	2016
Kevin Love: 2014	25	26.9	0.591	...	True	Kevin Love	2014
Russell Westbrook: 2015	26	29.1	0.536	...	True	Russell Westbrook	2015
James Harden: 2015	25	26.7	0.605	...	True	James Harden	2015



	Year	Player	Tm	...	PTS	PPG	ALL_STAR_PROB
6662	2017	James Harden	HOU	...	2356	29.086420	1.00
6915	2017	Russell Westbrook	OKC	...	2558	31.580247	1.00
6781	2017	LeBron James	CLE	...	1954	26.405405	1.00
6687	2017	Jimmy Butler	CHI	...	1816	23.894737	0.99
6749	2017	Kevin Durant	GSW	...	1555	25.080645	0.99
6741	2017	Kawhi Leonard	SAS	...	1888	25.513514	0.99
6457	2017	Anthony Davis	NOP	...	2099	27.986667	0.99
6644	2017	Isaiah Thomas	BOS	...	2199	28.934211	0.98
6941	2017	Stephen Curry	GSW	...	1999	25.303797	0.96
6540	2017	Damian Lillard	POR	...	2024	26.986667	0.95



What are your thoughts? I'm open to suggestions and/or more discussion.

Thanks,
Jake