Project: Indycar Rank Prediction
Progress Report
10282019-11082019

Abstract: The major progress in the past two weeks is building a stage dataset and prediction model on it according to the methods proposed in NASCAR race prediction[1][2]. On the task of predicting rank changes for each ‘stint’, a machine learning model does exceed the baselines.

Problem:

Due to the dynamics of a race, the ranks of the cars change abruptly in the laps with pitstops, as in Fig.1. The rank of a car then is correlated with its team’s strategy of pitstop and also other teams strategies. It is very complex to build a model to predict rank directly.

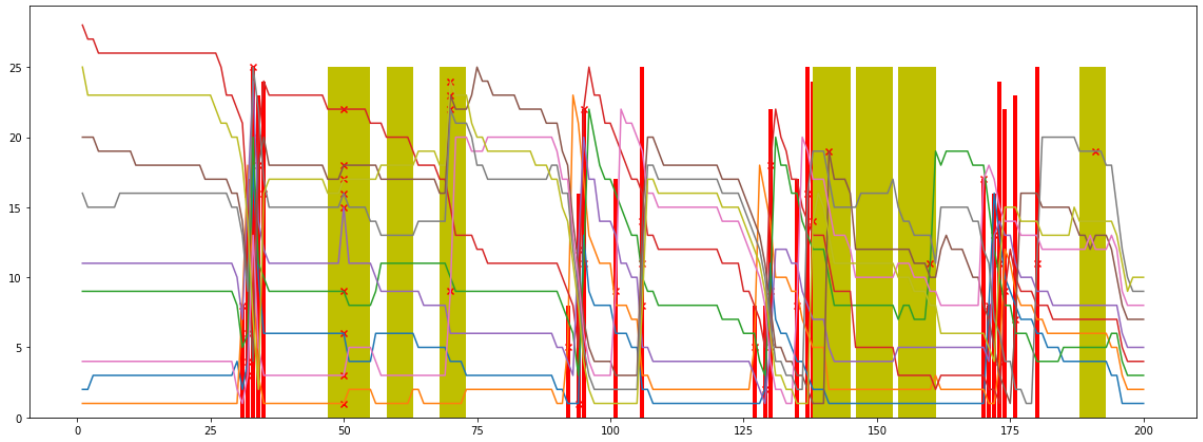

Fig.1 Ranks evolving along with the laps. Top 10 cars with the start position in Indy500. Red bar and the cross marker are pitstops. Yellow bars are caution laps.

Rank loss or gain in case of pitstop does not matter much. Because they are temporary. The car did pitstop will surpass the others who do pitstop later. It is not useful to predict accurately on these temporary rank values. Instead, paper[1][2] proposed to modeling the rank change in a larger scale, consecutive laps which start and end by a pitstop or a caution lap, as shown in the next Figure.

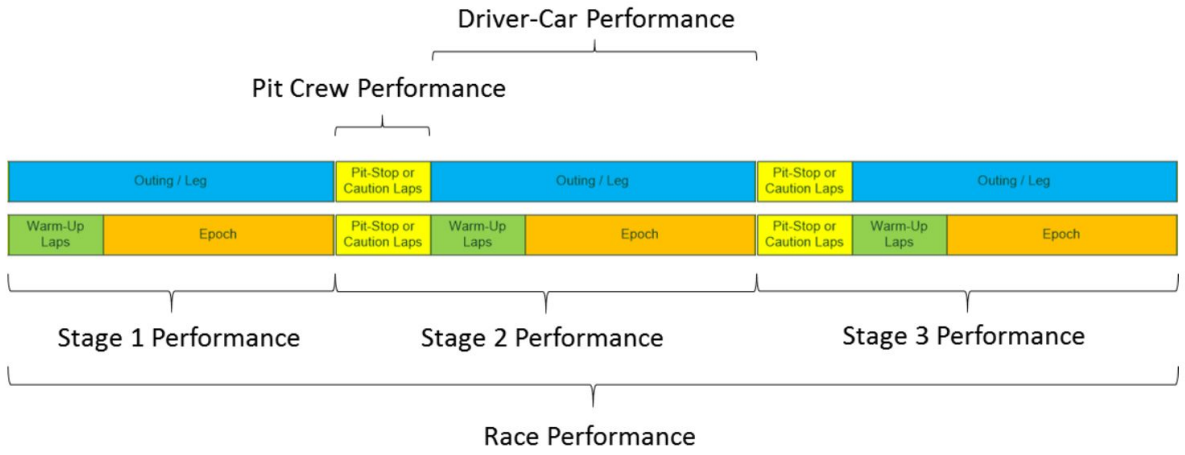

Figure 5. Composition of Race Performance based on Leg, Driver-Car, and Pit Crew based on a hypothetical Race comprising 3 Stages

In IndyCar, ‘stint’ refers to the period of driving between pit stops, similar to the term ‘stage’. We define our task to be the rank change prediction on stage dataset. A stage is defined as the laps between pit stops, adding a warmup period at the beginning denoted as ‘trim’. We do not add separate stages for caution laps as they don’t have many influences on rank.

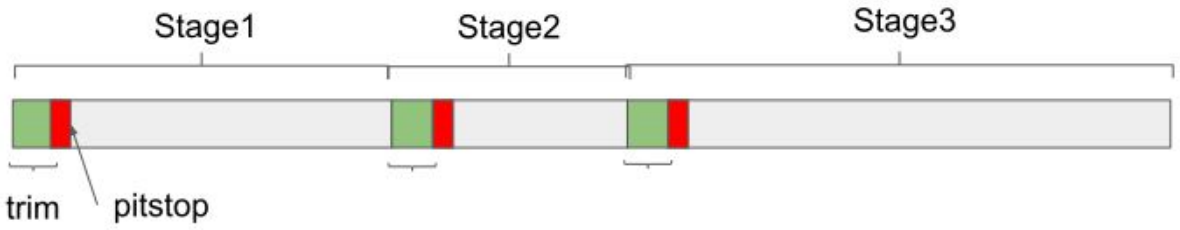

Fig.2 A race split into stages.

The task is to predict the rank change for each stage. When a car goes into the pit lanes, the model can predict how well the car will perform in the coming stage, whether it would lose or gain in rank. With the support of this prediction model, it potentially enables the team to evaluate different strategies, such as when to pit stop.

Dataset:

We have the timing and score log file of 2018 IndyCar series, which contains 17 events. Six of them are racing on the oval speedway, as Sfc=P in the following Table.

#	Date	Site	Cars	Winner(s)	St	C/E/T	Len	Sfc	Miles	Purse	Pole	Cau	Laps	Speed	LC
1	03/11/18	St. Petersburg	24	Sebastien Bourdais	14	DHF	1.800	S	198		105.085	8	25	86.207	11
2	04/07/18	Phoenix	23	Josef Newgarden	7	DCF	1.022	P	256		188.539	2	23	147.395	12
3	04/15/18	Long Beach	24	Alexander Rossi	1	DHF	1.968	S	167		106.454	4	17	88.622	6
4	04/23/18	Birmingham	23	Josef Newgarden	1	DCF	2.300	R	189		122.773	2	14	93.335	4
5	05/12/18	Indianapolis G.P.	24	Will Power	1	DCF	2.439	R	207		125.761	2	8	113.318	9
6	05/27/18	Indianapolis	33	Will Power	3	DCF	2.500	P	500		229.618	7	41	166.935	30
7	06/02/18	Belle Isle	23	Scott Dixon	2	DHF	2.350	S	165		113.024	2	10	99.285	6
8	06/03/18	Belle Isle	23	Ryan Hunter-Reay	10	DHF	2.350	S	165		90.661	1	3	105.176	6
9	06/09/18	Fort Worth	22	Scott Dixon	7	DHF	1.500	P	372		220.613	3	29	177.250	9
10	06/24/18	Elkhart Lake	23	Josef Newgarden	1	DCF	4.014	R	221		140.020	0	0	132.101	2
11	07/08/18	Iowa	22	James Hinchcliffe	11	DHF	0.894	P	268		182.391	2	17	149.737	4
12	07/15/18	Toronto	23	Scott Dixon	2	DHF	1.786	S	152		108.068	3	12	93.898	9
13	07/29/18	Mid-Ohio	24	Alexander Rossi	1	DHF	2.258	R	203		125.677	0	0	116.957	5
14	08/19/18	Pocono	22	Alexander Rossi	3	DHF	2.500	P	500		219.511	2	10	191.304	11
15	08/25/18	Gateway	21	Will Power	4	DCF	1.250	P	310		NTT	2	16	155.644	10
16	09/02/18	Portland	25	Takuma Sato	20	DHF	1.967	R	207		123.292	4	18	102.971	9
17	09/16/18	Sonoma	25	Ryan Hunter-Reay	1	DHF	2.385	R	203		110.605	1	5	99.440	5

Extract from these events, we get a stage dataset with 805 stage records. Features for each stage including the initial start position, start rank, change of rank, rate of changes and information of its neighbors. Detailed descriptions can be found in the Appendix.

Event	Records
Phoenix	114
Indy500	225
Texas	127
Iowa	109
Pocono	126
Gateway	104

Experiment Results:

First, we have two types of tasks: predict the value of rank change for a stage which is a regression problem, and predict the sign of rank change for a stage which is a classification problem.

We have three baselines:

CurRank	predict the end rank with the current start rank of the stage, i.e., means change is always zero.
AvgRank	predict the end rank with the average rank changes in previous stages.
Dice	predict the rank change by randomly throw dice which follows the distribution of the training data. For the classification task, it is a three facets dice(+,0,-).

We have two ways to split the dataset into training and test set.

split by event	select 5 events to train, the other 1 to test
split by stage	select the beginning stages to train, the left stages to test

Results of predicting the sign of rank change:

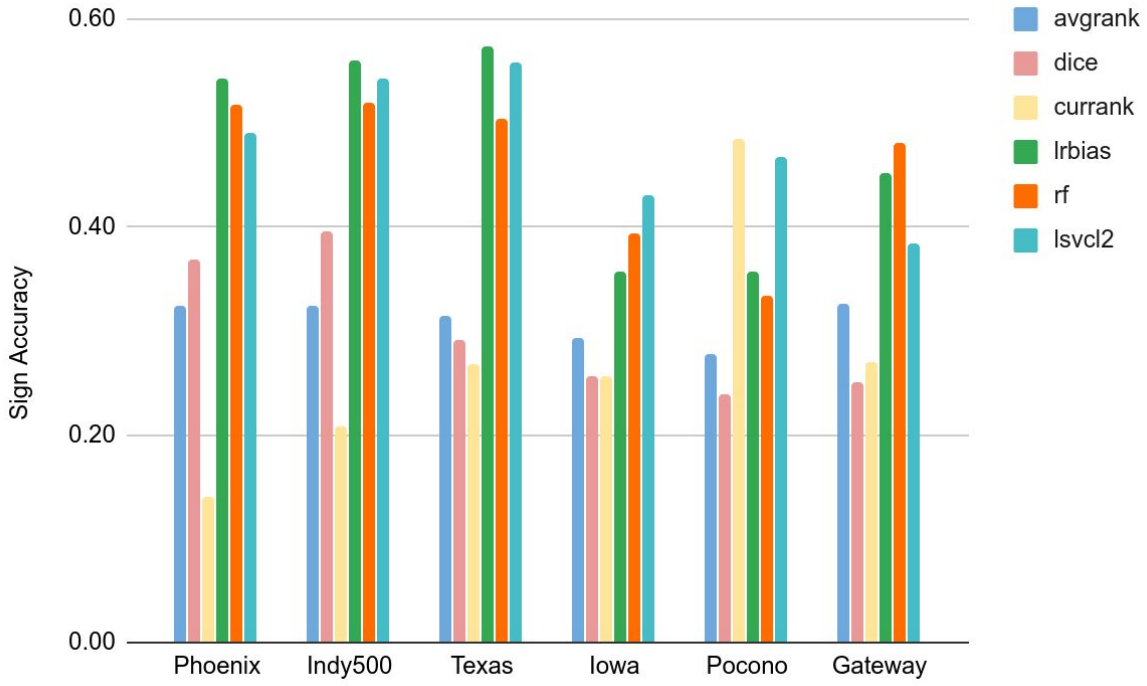

Fig.3 Sign of rank change prediction with split_by_events. Learning models are significantly better than the baselines, except one outlier currank is the best in Pocono, where more than no rank changes for 60% stages.

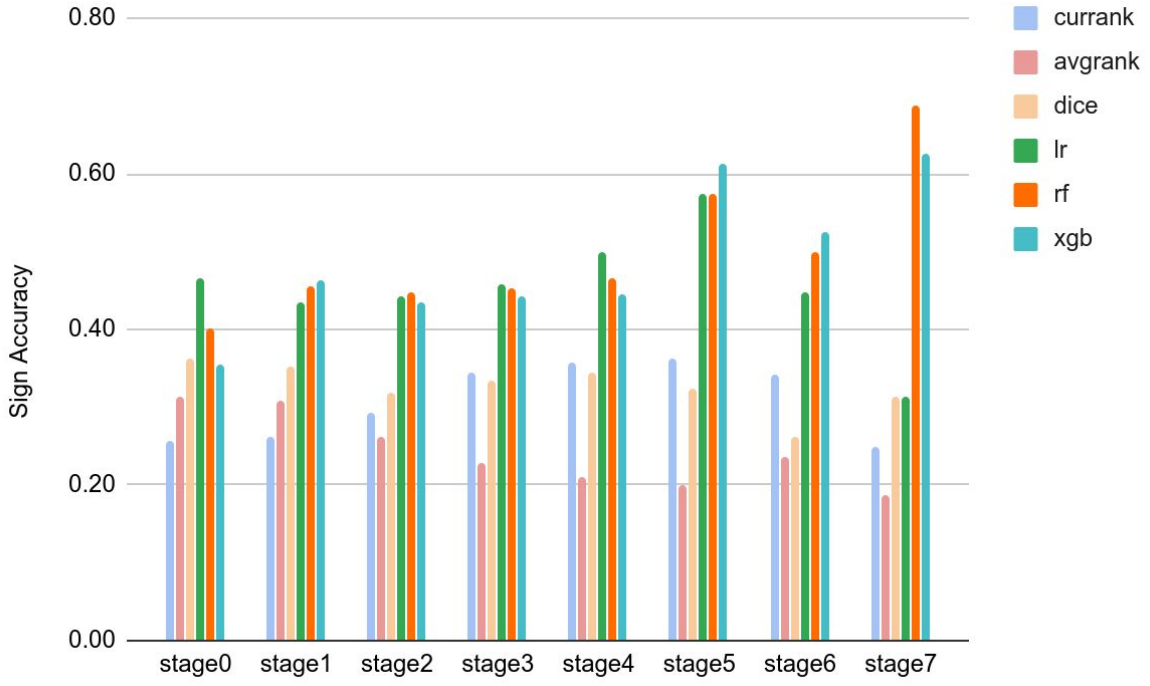

Fig.4 Sign of rank change prediction with split_by_stage. Learning models are significantly better than the baselines, and the gap increase when trained with more stages.

Results of predicting the value of rank change:

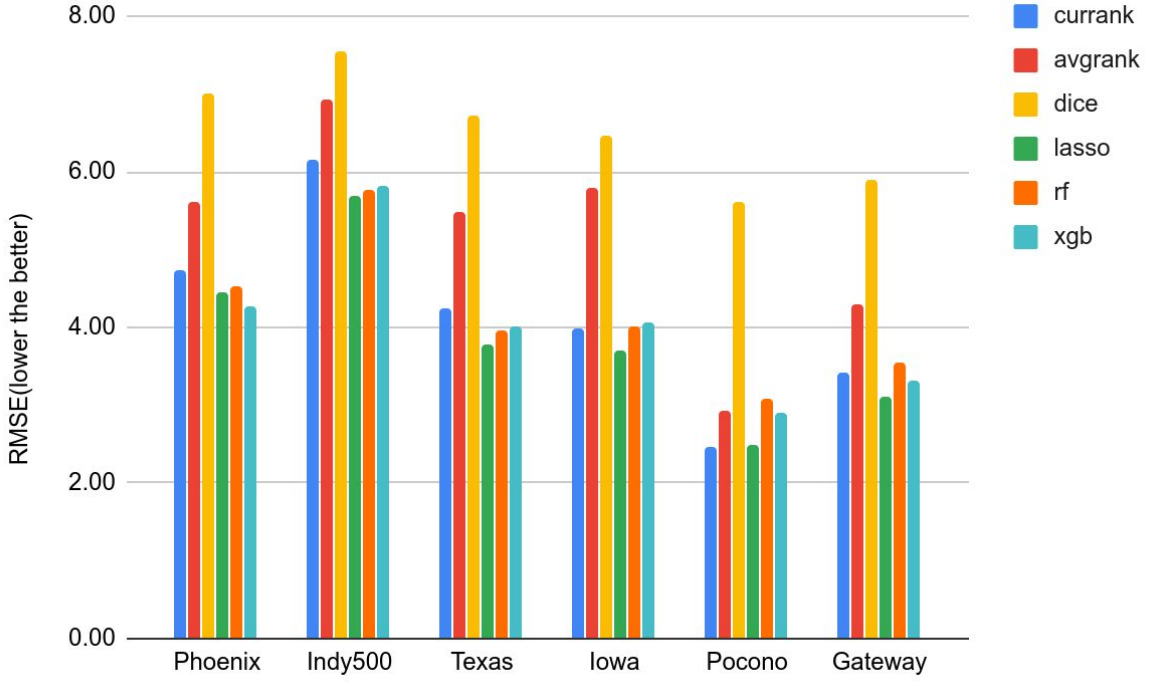

Fig.5 Rank change prediction with split_by_event. Learning models are only slightly better than the baselines.

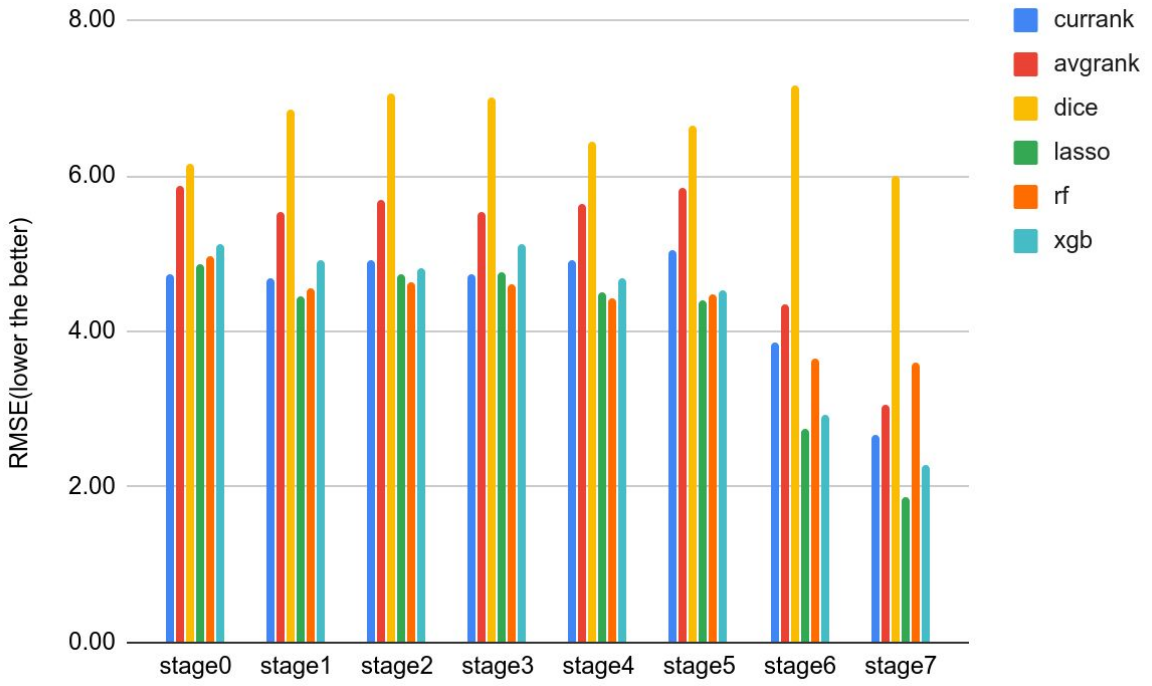

Fig.6 Rank change value prediction with split_by_stage. Learning models are slightly better than the baselines, and the gap increase when trained with more stages.

Conclusion:

On the rank change prediction task, model can be learned on stage dataset to achieve better performance than the baselines. This is consistent with the results reported by paper[1][2]. Sign prediction achieves similar performance, but the value predicted are not as good as theirs(Fig.5,Fig.6). There should be space to improve, e.g., by incorporating more features.

Work TODO:

1. Most of importance. Communicate with IndyCar to have a discussion on whether the result of rank change prediction is useful to them? or potentially useful to them.
2. Need help to verify the code and results are correct.
3. Continue to improve the model, by adding more features.
4. Think about other prediction tasks. Because the dataset is still limited in its size. Considering to add extra information about the drivers, such as text, news, etc, could be another direction promising to explore.

References
[1]T. Tulabandhula and C. Rudin, “Tire changes, fresh air, and yellow flags: challenges in predictive analytics for professional racing,” Big data, vol. 2, no. 2, pp. 97–112, 2014.
[2]C. L. W. Choo, “Real-time decision making in motorsports: analytics for improving professional car race strategy,” PhD Thesis, Massachusetts Institute of Technology, 2015.

Appendix:

1. **Features**

Type	Feature	Meaning
gobal info	stageid	
	firststage	1/0
	pit_in_caution	1/0
	start_position	#
0 order of #rank	start_rank	#rank
	start_rank_ratio	#rank/carnum
	top_pack	top5 1/0
	bottom_pack	bottom5 1/0
	average_rank	previous stage
	average_rank_all	all previous stages
1 order of #rank	change_in_rank	previous stage
	change_in_rank_all	all previous stages
2 order of #rank	rate_of_change	previous stage
	rate_of_change_all	all previous stages
neighbors	prev_nb0_change_in_rank	previous car
	prev_nb1_change_in_rank	
	prev_nb2_change_in_rank	
	follow_nb0_change_in_rank	following car
	follow_nb1_change_in_rank	
	follow_nb2_change_in_rank	

2. **Results**

Table: for Fig.3 Sign of rank change prediction with split_by_events.

event	trainsi	testsi	testdi	curra	avgra	dice	lr	lrl1	lsvc	lsvcl2	rf	lrbias	xgb

	ze	ze	stribut ion	nk	nk								
Phoe nix	691	114	+:38, 0:16,- :60	0.14	0.32	0.37	0.54	0.55	0.50	0.49	0.52	0.54	0.44
Indy5 00	580	225	+:82, 0:47,- :96	0.21	0.32	0.40	0.56	0.56	0.56	0.54	0.52	0.56	0.53
Texas	678	127	+:39, 0:34,- :54	0.27	0.31	0.29	0.56	0.57	0.60	0.56	0.50	0.57	0.45
Iowa	696	109	+:42, 0:28,- :39	0.26	0.29	0.26	0.35	0.36	0.36	0.43	0.39	0.36	0.30
Poco no	679	126	+:29, 0:61,- :36	0.48	0.28	0.24	0.37	0.37	0.35	0.47	0.33	0.36	0.46
Gate way	701	104	+:34, 0:28,- :42	0.27	0.33	0.25	0.48	0.43	0.43	0.38	0.48	0.45	0.40

Table: for Fig.4 Sign of rank change prediction with split_by_stages.

runid	trainsi ze	testsi ze	testdi stribut ion	curra nk	avgra nk	dice	lr	lrl1	lsvc	lsvcl2	rf	lrbias	xgb
stage 0	153	652	+:221 ,0:16 7,-:26 4	0.26	0.31	0.36	0.46	0.45	0.44	0.27	0.40	0.45	0.36
stage 1	288	517	+:186 ,0:13 6,-:19 5	0.26	0.31	0.35	0.44	0.45	0.46	0.45	0.45	0.43	0.46
stage 2	421	384	+:140 ,0:11 2,-:13 2	0.29	0.26	0.32	0.44	0.43	0.46	0.42	0.45	0.44	0.43
stage 3	547	258	+:91, 0:89,- :78	0.34	0.23	0.33	0.46	0.44	0.45	0.41	0.45	0.46	0.44
stage 4	657	148	+:48, 0:53,- :47	0.36	0.21	0.34	0.50	0.50	0.47	0.43	0.47	0.51	0.45
stage	725	80	+:26,	0.36	0.20	0.33	0.58	0.59	0.55	0.54	0.58	0.59	0.61

5			0:29,-:25										
stage 6	767	38	+:11, 0:13,-:14	0.34	0.24	0.26	0.45	0.45	0.47	0.37	0.50	0.47	0.53
stage 7	789	16	+:4,0:4,-:8	0.25	0.19	0.31	0.31	0.38	0.31	0.25	0.69	0.31	0.63

Table: for Fig.5 Rank change prediction with split_by_event.

runid	trainsize	testsize	testdistribution	currank	avgrank	dice	lasso	ridge	rf	svr	xgb
Phoenix	691	114	+:38,0:16,-:60	4.73	5.61	7.01	4.44	4.42	4.53	4.70	4.26
Indy500	580	225	+:82,0:47,-:96	6.17	6.94	7.55	5.69	5.51	5.76	6.14	5.82
Texas	678	127	+:39,0:34,-:54	4.26	5.48	6.73	3.78	3.76	3.96	4.13	4.02
Iowa	696	109	+:42,0:28,-:39	3.98	5.80	6.45	3.69	3.86	4.02	3.91	4.06
Pocono	679	126	+:29,0:61,-:36	2.48	2.93	5.61	2.49	2.57	3.07	2.38	2.91
Gateway	701	104	+:34,0:28,-:42	3.41	4.30	5.89	3.11	3.11	3.54	3.34	3.32

Table: for Fig.6 Rank change value prediction with split_by_stage.

runid	trainsize	testsize	testdistribution	currank	avgrank	dice	lasso	ridge	rf	svr	xgb
stage0	153	652	+:221, 0:167,-:264	4.75	5.87	6.16	4.85	4.96	4.98	4.76	5.13
stage1	288	517	+:186, 0:136,-:195	4.68	5.53	6.86	4.45	4.69	4.55	4.78	4.92
stage2	421	384	+:140, 0:112,-:132	4.90	5.68	7.05	4.74	4.62	4.64	5.00	4.80
stage3	547	258	+:91,0:89,-:78	4.73	5.54	7.01	4.76	4.87	4.62	4.82	5.12
stage4	657	148	+:48,0:53,-:47	4.91	5.65	6.45	4.51	4.56	4.43	4.95	4.69

stage5	725	80	+:26,0:29,-:25	5.05	5.85	6.65	4.39	4.37	4.49	5.08	4.52
stage6	767	38	+:11,0:13,-:14	3.85	4.36	7.17	2.74	2.68	3.65	3.86	2.92
stage7	789	16	+:4,0:4,-:8	2.68	3.05	5.99	1.87	1.77	3.60	2.67	2.28