NFL Game Outcome Prediction

Hardy Leung, MS-AI (kwok-shing.leung@sjsu.edu) [SJSU ID: 016711877]

The National Football League

- The true American pastime
- NFL Founded in 1920
- Football *is* America
- Football is big business
    - 24% sports betting
    - Fantasy sports $24B in 2022

The Beal Paper

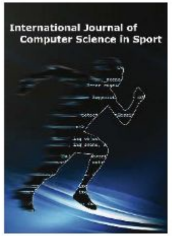
International Journal of Computer Science in Sport

Volume 19, Issue 2, 2020

Journal homepage: http://iacss.org/index.php?id=30

sciendo

DOI: 10.2478/ijcss-2020-0009



A Critical Comparison of Machine Learning Classifiers to Predict Match Outcomes in the NFL

Ryan Beal, Timothy J. Norman and Sarvapali D. Ramchurn

University of Southampton, University Rd, Southampton, SO17 1BJ

Abstract

In this paper, we critically evaluate the performance of nine machine learning classification techniques when applied to the match outcome prediction problem presented by American Football. Specifically, we implement and test nine techniques using real-world datasets of 1280 games over 5 seasons from the National Football League (NFL). We test the nine different classifier techniques using a total of 42 features for each team and we find that the best performing algorithms are able to improve one previous published works. The algoriothms achieve an accuracy of between 44.64% for a Guassian Process classifier to 67.53% with a Naïve Bayes classifer. We also test each classifier on a year by year basis and compare our results to those of the bookmakers and other leading academic papers.

KEYWORDS: MACHINE LEARNING, SUPERVISED LEARNING, FOOTBAL, NFL

Beal's Models

Support Vector Machine Decision Tree Naïve Bayes	Nearest Neighbors Random Forest Quadratic Discriminant Analysis (QDA)	Gaussian Process AdaBoost Neural Network

- They addressed the problem of NFL game prediction head on, using actual datasets curated from actual games
- They surveyed nine different machine learning techniques, which offer a good overview of how these methods performed in a complex benchmark

The Beal Paper

Setup

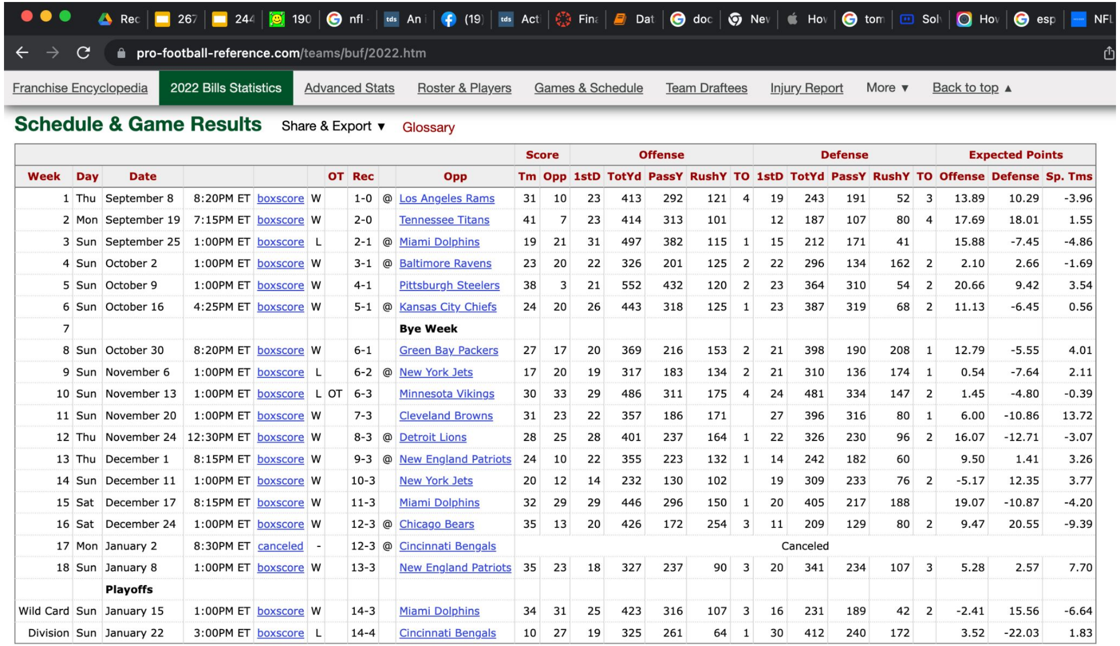
- Dataset = the full-game statistical summary of 1280 games played by all 32 NFL teams in a five-year period
- Make W/L predictions
- A simplistic model with limited feature selections
  - 21 distinct concepts such as passing completions, total yards, etc,
  - Home + Away, This season + Last season
  - 21 x 4 + 1 = 85

Significant Opportunities

- Bigger Dataset (2006 – 2023)
- Better Statistics
- ESPN QBR Rating
- 538 ELO Rating

Pro-Football-Reference.com

- Beal’s Data Source
- Web scrapping



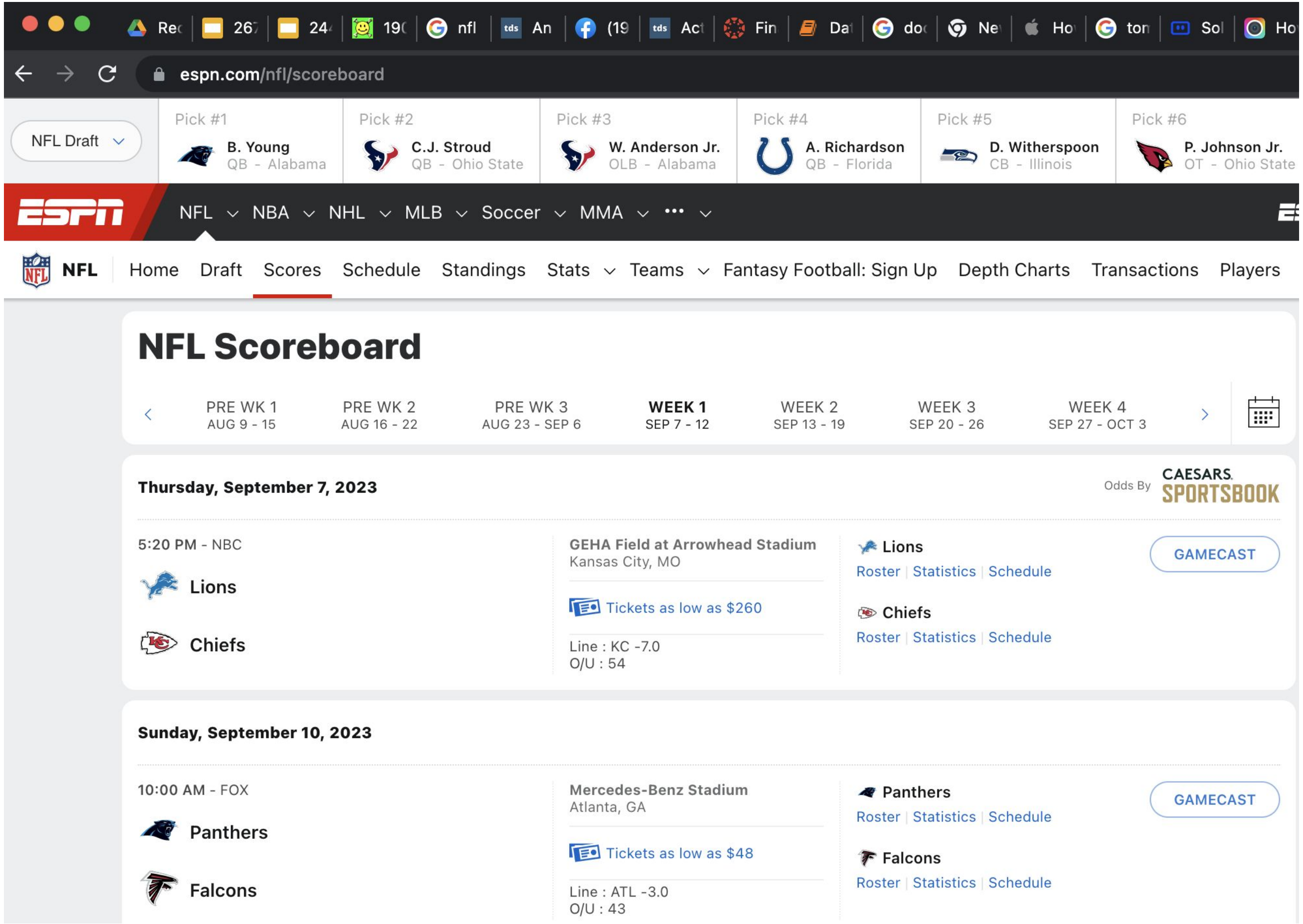
Schedule & Game Results

Share & Export ▼ Glossary

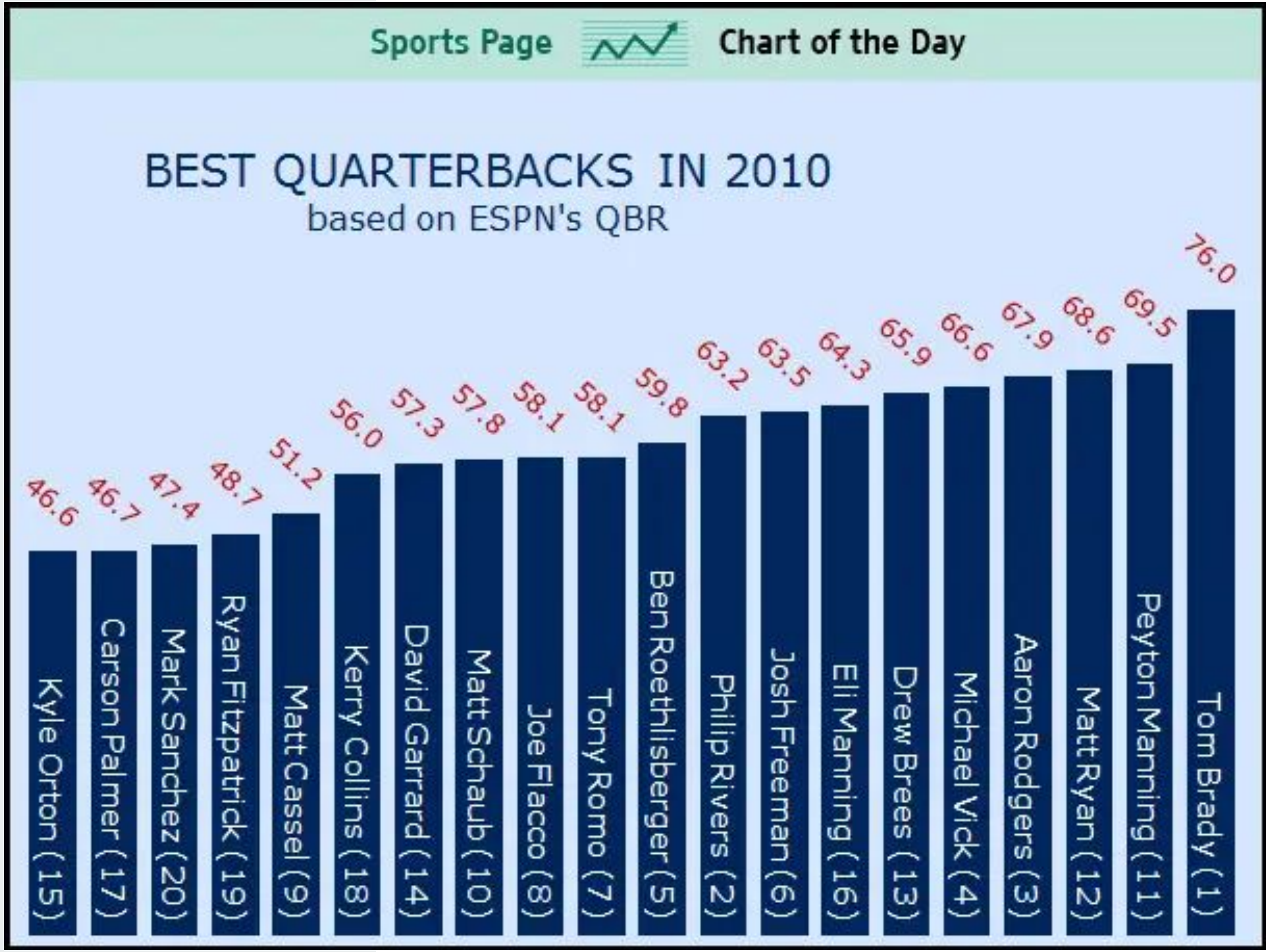
										Score		Offense					Defense					Expected Points		
Week	Day	Date				OT	Rec		Opp	Tm	Opp	1stD	TotYd	PassY	RushY	TO	1stD	TotYd	PassY	RushY	TO	Offense	Defense	Sp. Tms
1	Thu	September 8	8:20PM ET	boxscore	W		1-0	@	Los Angeles Rams	31	10	23	413	292	121	4	19	243	191	52	3	13.89	10.29	-3.96
2	Mon	September 19	7:15PM ET	boxscore	W		2-0		Tennessee Titans	41	7	23	414	313	101		12	187	107	80	4	17.69	18.01	1.55
3	Sun	September 25	1:00PM ET	boxscore	L		2-1	@	Miami Dolphins	19	21	31	497	382	115	1	15	212	171	41		15.88	-7.45	-4.86
4	Sun	October 2	1:00PM ET	boxscore	W		3-1	@	Baltimore Ravens	23	20	22	326	201	125	2	22	296	134	162	2	2.10	2.66	-1.69
5	Sun	October 9	1:00PM ET	boxscore	W		4-1		Pittsburgh Steelers	38	3	21	552	432	120	2	23	364	310	54	2	20.66	9.42	3.54
6	Sun	October 16	4:25PM ET	boxscore	W		5-1	@	Kansas City Chiefs	24	20	26	443	318	125	1	23	387	319	68	2	11.13	-6.45	0.56
7									**Bye Week**															
8	Sun	October 30	8:20PM ET	boxscore	W		6-1		Green Bay Packers	27	17	20	369	216	153	2	21	398	190	208	1	12.79	-5.55	4.01
9	Sun	November 6	1:00PM ET	boxscore	L		6-2	@	New York Jets	17	20	19	317	183	134	2	21	310	136	174	1	0.54	-7.64	2.11
10	Sun	November 13	1:00PM ET	boxscore	L	OT	6-3		Minnesota Vikings	30	33	29	486	311	175	4	24	481	334	147	2	1.45	-4.80	-0.39
11	Sun	November 20	1:00PM ET	boxscore	W		7-3		Cleveland Browns	31	23	22	357	186	171		27	396	316	80	1	6.00	-10.86	13.72
12	Thu	November 24	12:30PM ET	boxscore	W		8-3	@	Detroit Lions	28	25	28	401	237	164	1	22	326	230	96	2	16.07	-12.71	-3.07
13	Thu	December 1	8:15PM ET	boxscore	W		9-3	@	New England Patriots	24	10	22	355	223	132	1	14	242	182	60		9.50	1.41	3.26
14	Sun	December 11	1:00PM ET	boxscore	W		10-3		New York Jets	20	12	14	232	130	102		19	309	233	76	2	-5.17	12.35	3.77
15	Sat	December 17	8:15PM ET	boxscore	W		11-3		Miami Dolphins	32	29	29	446	296	150	1	20	405	217	188		19.07	-10.87	-4.20
16	Sat	December 24	1:00PM ET	boxscore	W		12-3	@	Chicago Bears	35	13	20	426	172	254	3	11	209	129	80	2	9.47	20.55	-9.39
17	Mon	January 2	8:30PM ET	canceled	-		12-3	@	Cincinnati Bengals	Canceled														
18	Sun	January 8	1:00PM ET	boxscore	W		13-3		New England Patriots	35	23	18	327	237	90	3	20	341	234	107	3	5.28	2.57	7.70
		Playoffs																						
Wild Card	Sun	January 15	1:00PM ET	boxscore	W		14-3		Miami Dolphins	34	31	25	423	316	107	3	16	231	189	42	2	-2.41	15.56	-6.64
Division	Sun	January 22	3:00PM ET	boxscore	L		14-4		Cincinnati Bengals	10	27	19	325	261	64	1	30	412	240	172		3.52	-22.03	1.83

ESPN

- Model
- Similar Statistics
- JSON APIs



ESPN QBR



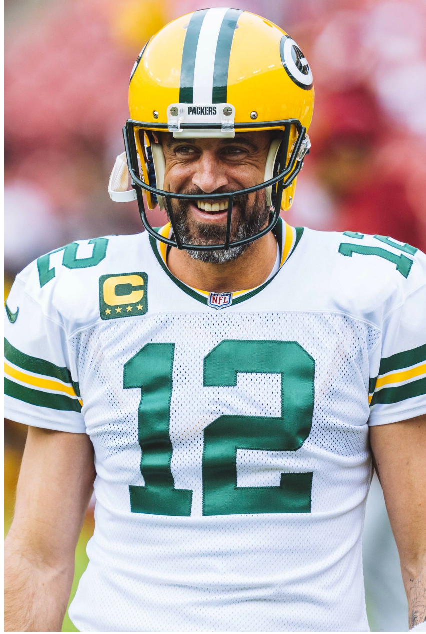
Sports Page
Chart of the Day
BEST QUARTERBACKS IN 2010
based on ESPN's QBR
46.6
46.7
47.4
48.7
51.2
56.0
57.3
57.8
58.1
58.1
59.8
63.2
63.5
64.3
65.9
66.6
67.9
68.6
69.5
76.0
Kyle Orton (15)
Carson Palmer (17)
Mark Sanchez (20)
Ryan Fitzpatrick (19)
Matt Cassel (9)
Kerry Collins (18)
David Garrard (14)
Matt Schaub (10)
Joe Flacco (8)
Tony Romo (7)
Ben Roethlisberger (5)
Philip Rivers (2)
Josh Freeman (6)
Eli Manning (16)
Drew Brees (13)
Michael Vick (4)
Aaron Rodgers (3)
Matt Ryan (12)
Peyton Manning (11)
Tom Brady (1)

ESPN QBR

“QBR incorporates all of a quarterback’s contributions to winning, including how he impacts the game on passes, rushes, turnovers and penalties. Also, since QBR is built from the play level, it accounts for a team’s level of success or failure on every play to provide the proper context and then allocates credit to the quarterback and his teammate to produce a clearer measure of quarterback efficiency”

– Katz and Burke, ESPN

ESPN QBR



Caveats

- ESPN data can be missing
- ESPN data can be messy (person != team)
- ESPN data can be WRONG
  - Aaron Rodgers did not play for Jets for the last two years!

538's ELO Rating

LAST EDIT JAN. 9, 2023

How Our NFL Predictions Work

Filed under Methodology

See our latest predictions

Pro-Football-Reference.com

Autocorrelation / Elo rating / Monte Carlo simulations / Regression to the mean / ESPN's Total Quarterback Rating

FiveThirtyEight has an admitted fondness for the Elo rating — a simple system that judges teams or players based on head-to-head results — and we've used it to rate competitors in basketball, baseball, tennis and various other sports over the years. The sport we cut our teeth on, though, was professional football. Way back in 2014, we developed our NFL Elo ratings to forecast the outcome of every game. The nuts and bolts of that system are described below.

Game predictions

In essence, Elo assigns every team a power rating (the NFL average is around 1500). Those ratings are then used to generate win probabilities for games, based on the difference in quality between the two teams involved, plus adjustments for changes at starting quarterback, the location of the

Data Processing – 11 Steps

1. Read ELO (clean, backbone)
2. Read ESPN (via Kaggle dataset)
3. Create game dictionary (team, date)
4. Read ESPN QBR
5. Calendar management (2023/02/12 = which week? And vice versa)
6. Fix ESPN QBR problem
7. Process ELO
8. Split game into two halves
9. Compute averages of the week
10. Compute rolling averages (multiple, based on SPECs)
11. For each (rivalry, date)
    - Pull necessary averages based on SPECs

Data Processing – Why So Complicated?

- Try out different ways of extracting the features while maintaining correctness
- Avoid data leakage
- Flexible and extensible for prosperity

Tabular Data

- 3 sets of averages
  - Last game, this season, last season
  - 33 features / team
  - 5 + 33 x 2 x 3 = 5 + 198 = 203
- Each model in the experiment takes the exact same tabular data
  - 4573 rows, 203 columns

Experimental Results

Algorithm	Beal et al. [2]	Our Work
SVM with RBF	Yes	Yes
Nearest Neighbors	Yes	Yes ($\star$)
Gaussian Process	Yes	Yes
Decision Tree	Yes	Yes
Random Forest	Yes	Yes
AdaBoost	Yes	Yes
Naïve Bayes	Yes	Yes
QDA	Yes	Yes
Neural Network	Yes	Yes ($\star$)
Logistic Regression	Yes	Yes
XGBoost		Yes
LightGBM		Yes
CatBoost		Yes
Model Tree		Yes
TabNet		Yes
Ensemble		Yes
Elo		Yes

Exp I

Algorithm	Rank	Accuracy	Precision	Recall	F1
Gaussian Process	16	0.4706	–	–	–
QDA	15	0.5294	0.5294	1.0000	0.6923
TabNet	14	0.5376	0.5564	0.6239	0.5882
Decision Tree	13	0.5756	0.5929	0.6325	0.6121
XGBoost	12	0.5864	0.5924	0.7009	0.6421
SVM	11	0.6109	0.5878	0.8872	0.7071
AdaBoost	10	0.5937	0.6009	0.6923	0.6434
Naïve Bayes	9	0.5973	0.5875	0.8034	0.6787
LightGBM	8	0.6045	0.6034	0.7385	0.6641
Ensemble (Mult)	7	0.6054	0.5959	0.7915	0.6799
Ensemble (Sum)	6	0.6072	0.5984	0.7846	0.6790
Ensemble (Vote)	5	0.6136	0.6065	0.7692	0.6782
Random Forest	4	0.6109	0.6078	0.7470	0.6702
Model Tree	3	0.6163	0.6148	0.7368	0.6703
Logistic Regression	2	0.6325	0.6211	0.7846	0.6934
Elo	1	0.6470	0.6696	0.6581	0.6638

Table 7: Experimental I: $\alpha = 1.00$, $\epsilon = 1.00$, Unnormalized, No PCA.

Logistic Regression

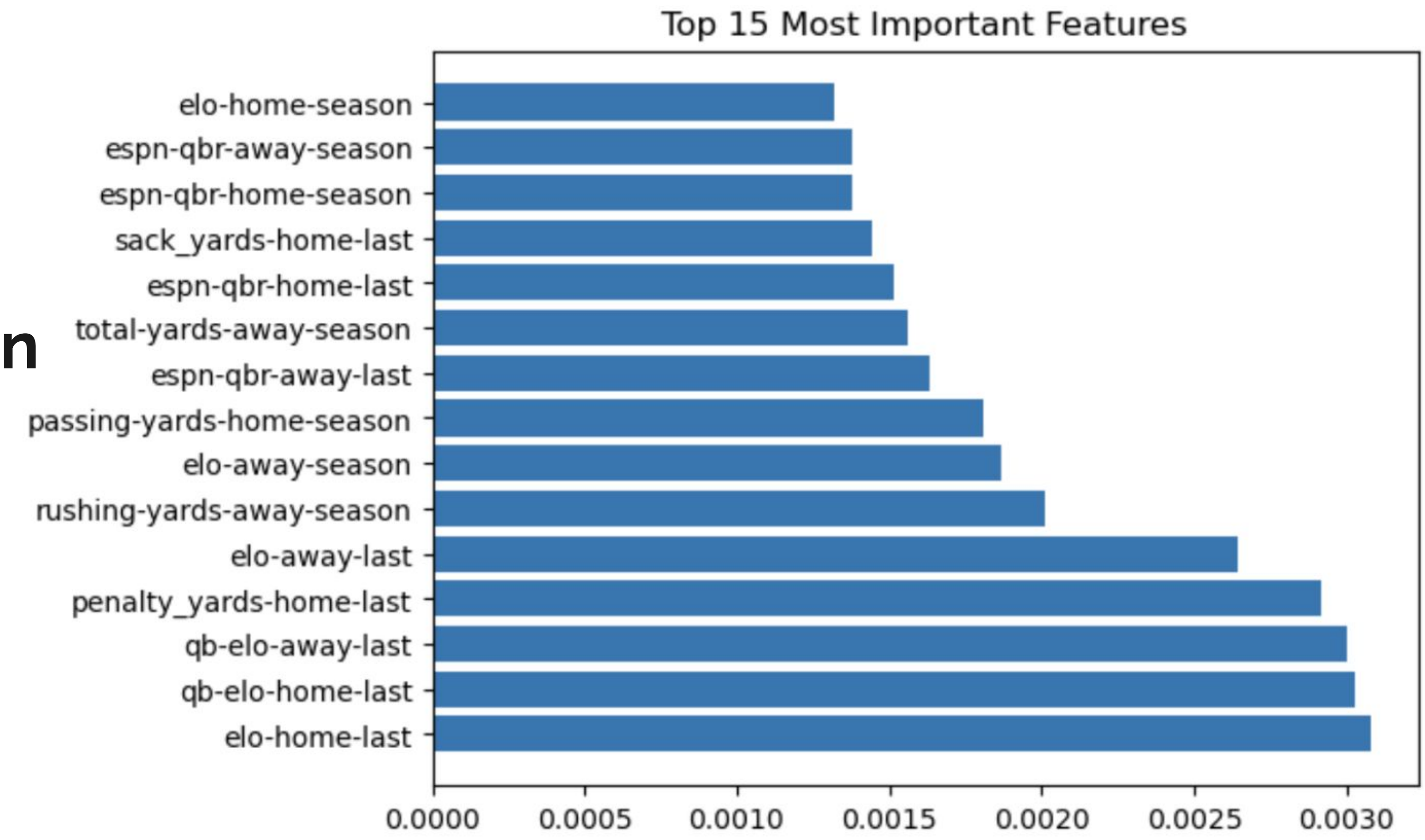


Figure 2: The top-15 most importance features.

Exp II

+PCA

Algorithm	Rank	Accuracy	Precision	Recall	F1
Gaussian Process	16	0.4706	–	–	–
TabNet	15	0.5357	0.5502	0.6735	0.6057
Decision Tree	14	0.5475	0.5687	0.6017	0.5847
QDA	13	0.6009	0.5891	0.8137	0.6834
Random Forest	12	0.6072	0.6024	0.7590	0.6717
XGBoost	11	0.6090	0.6147	0.7009	0.6550
SVM	10	0.6100	0.5925	0.8427	0.6958
Model Tree	9	0.6163	0.6148	0.7368	0.6703
AdaBoost	8	0.6200	0.6132	0.7641	0.6804
LightGBM	7	0.6217	0.6223	0.7265	0.6703
Logistic Regression	6	0.6235	0.6125	0.7863	0.6886
Ensemble (Vote)	5	0.6271	0.6183	0.7726	0.6869
Naïve Bayes	4	0.6290	0.6061	0.8547	0.7092
Ensemble (Sum)	3	0.6299	0.6219	0.7675	0.6871
Ensemble (Mult)	2	0.6299	0.6222	0.7658	0.6866
Elo	1	0.6470	0.6696	0.6581	0.6638

Table 8: Experimental II: $\alpha = 1.00$, $\epsilon = 1.00$, Unnormalized, with PCA.

Exp III

+Normalization

Algorithm	Rank	Accuracy	Precision	Recall	F1
QDA	16	0.4823	0.5243	0.2393	0.3286
Naïve Bayes	15	0.4842	0.5333	0.2051	0.2963
Gaussian Process	14	0.5294	0.5294	1.0000	0.6923
Logistic Regression	13	0.5294	0.5294	1.0000	0.6923
Decision Tree	12	0.5303	0.5432	0.7094	0.6153
XGBoost	11	0.5665	0.5673	0.7641	0.6511
SVM	10	0.5709	0.5709	1.0000	0.7269
Model Tree	9	0.5744	0.5735	0.9929	0.7271
LightGBM	8	0.5756	0.6086	0.5556	0.5809
Random Forest	7	0.5765	0.5990	0.6051	0.6020
Ensemble (Vote)	6	0.5819	0.5718	0.8376	0.6796
TabNet	5	0.5919	0.5876	0.7675	0.6657
Ensemble (Mult)	4	0.5919	0.5840	0.7966	0.6739
AdaBoost	3	0.5937	0.5852	0.7983	0.6753
Ensemble (Sum)	2	0.5937	0.5852	0.7983	0.6753
Elo	1	0.6470	0.6696	0.6581	0.6638

Table 9: Experimental III: $\alpha = 1.00$, $\epsilon = 1.00$, Normalized, with PCA.

Exp IV

+Weight Decay

Algorithm	Rank	Accuracy	Precision	Recall	F1
QDA	16	0.4697	0.4935	0.0650	0.1148
Gaussian Process	15	0.4706	–	–	–
TabNet	14	0.5403	0.5399	0.8906	0.6723
Decision Tree	13	0.5692	0.5932	0.5932	0.5932
AdaBoost	12	0.5954	0.6117	0.6462	0.6284
Naïve Bayes	11	0.6018	0.5926	0.7932	0.6784
XGBoost	10	0.6027	0.6127	0.6786	0.6440
LightGBM	9	0.6063	0.6130	0.6957	0.6517
SVM	8	0.6090	0.5870	0.8821	0.7049
Ensemble (Vote)	7	0.6136	0.6079	0.7607	0.6758
Ensemble (Mult)	6	0.6235	0.6125	0.7863	0.6886
Ensemble (Sum)	5	0.6235	0.6125	0.7863	0.6886
Random Forest	4	0.6235	0.6181	0.7556	0.6780
Model Tree	3	0.6262	0.6229	0.7453	0.6786
Logistic Regression	2	0.6271	0.6151	0.7897	0.6916
Elo	1	0.6470	0.6696	0.6581	0.6638

Table 10: Experimental IV: $\alpha = 0.95$, $\epsilon = 0.50$, Unnormalized, No PCA.

Conclusion

Extended Beal's Work in Several Ways

- Used Bigger Dataset (350%)
- More Key statistics, ESPN Data, QBR, ELO
- Evaluated More Models
- Robust Data Processing Flow
- Ensemble Model
- Studied the impact of normalization, PCA, and weight decay

Much more research opportunities!