Predicting Season Long NFL Statistics

Company: ESPN

Problem Statement: ESPN is one of the largest providers of fantasy football services in the industry. Prior to each season they provide their users with season long predictions for each player. They want to see if I can build a better model than their current lead fantasy data scientist, Mike Clay.

Data Gathering: Initial Statistics

4 main positions: Quarterback (QB), Running Back (RB), Wide Receiver (WR), Tight End (TE)

Scraped NFL Player statistics (yards, TDs, yards per reception, etc.) from 2014 to 2018 (5 full seasons) from pro-football-reference.com

Not too much cleaning required. Normalized names by removing punctuation, made sure team names were consistent because a team changed its name in both 2016 and 2017.

Why not more data? The baseline stats are easy to scrape but it’s increasingly difficult to get the other information (coaching/coordinator changes, salary information) for previous years

Data Gathering: Target Column

Choose to model for fantasy points because it's just a combination of all a player's statistics in a single number

I took the top 300 fantasy scoring players from each year from 2014 to 2018 (5 full seasons) to model against.

It's cheating a little bit because we don't know who the top 300 players will be at the beginning of the year...

But it helps avoid some major hurdles such as having to adjust for players no longer in the league (retirement, suspension, etc) and I couldn't find a way to get that data without doing it by hand.

Data Gathering: Additional Features

I found websites with data on coaching and coordinator changes but it wasn't scrapable so I had to bring over that info by hand...

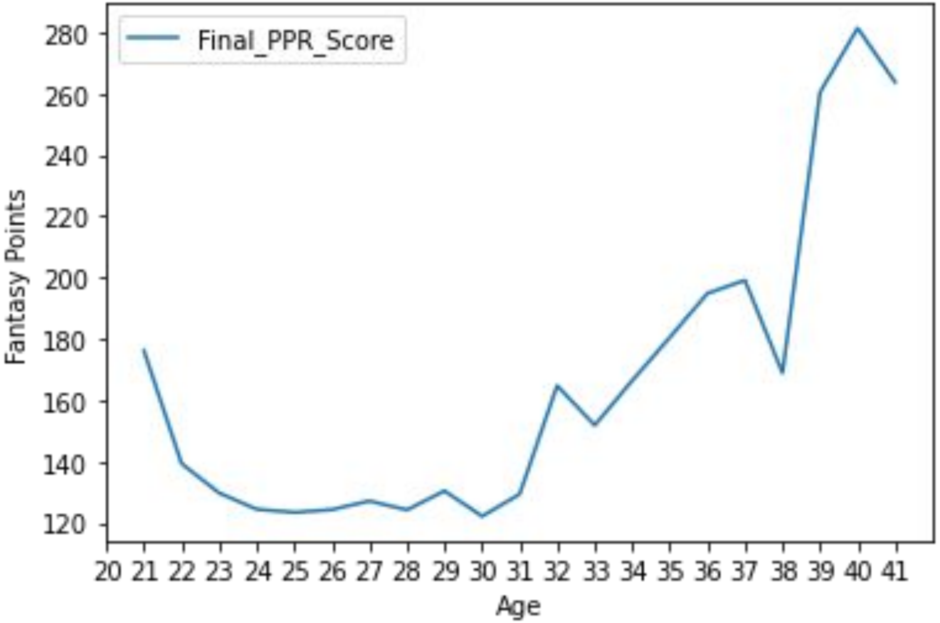
Aggregate preseason fantasy rankings by hundreds of "experts" pulled from FantasyPros.com

Salary data for each player from overthecap.com

Target data, i.e. the number of times a quarterback throws to a player whether he catches it or not. Known to be predictive in fantasy football so I wanted to get players' targets per game instead of total targets. This would adjust for any injuries or games missed. Scraped from footballguys.com

Wanted to have a contract year column (1 or 0 if a player was in the last year of their contract) but after further research on the subject, there didn't seem to be a correlation between that and performance (https://www.footballoutsiders.com/stat-analysis/2014/contract-year-phenomenon-revisited)

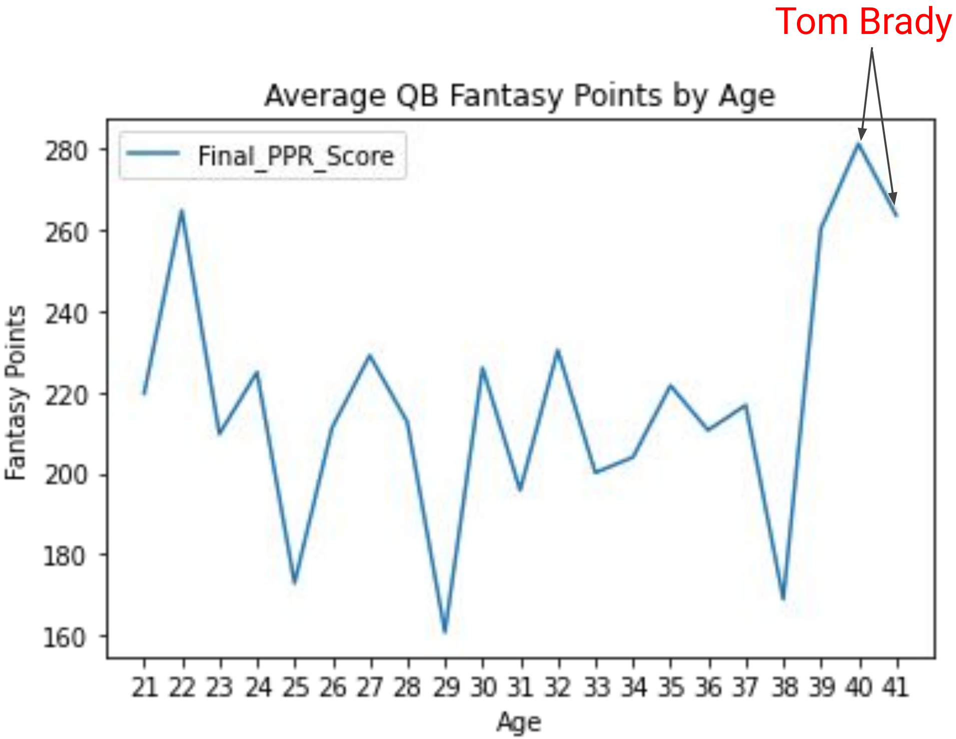
EDA: Age...Not What You Expect



We would expect fantasy points to decrease as age increases but the graph is thrown off by outliers in each positional category (QB, RB, WR, TE)

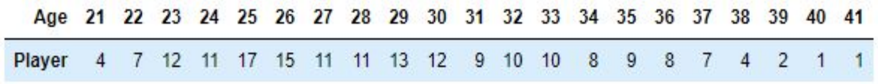
My hypothesis is that players playing in their 30s are limited to only really good players. So the data becomes skewed because there are less observations with high fantasy point scorers.

EDA: QB Age


Tom Brady

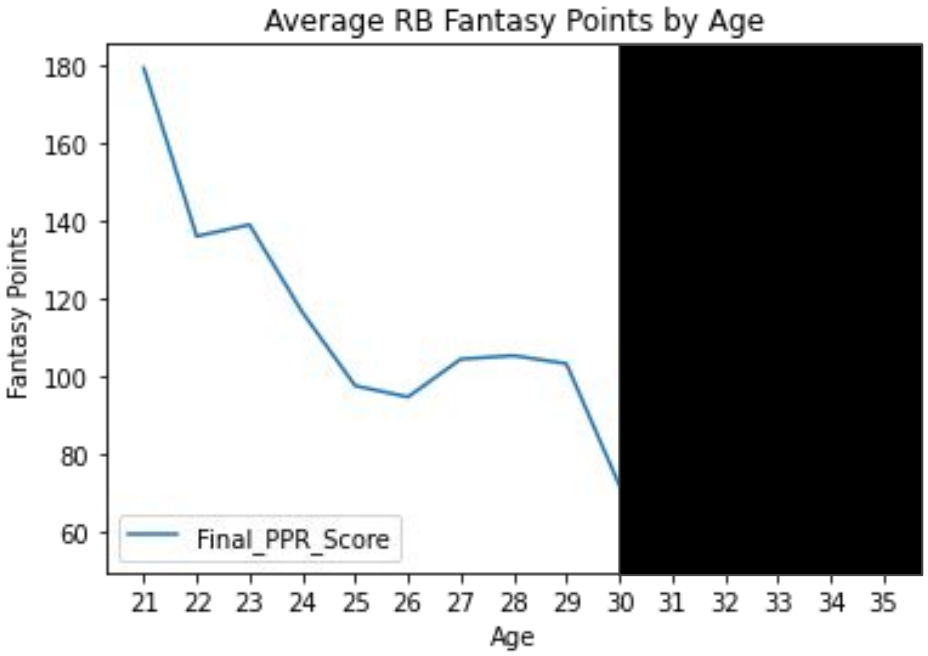
We don’t have a ton of QB observations so this graph is really distorted by outliers like Tom Brady who is playing really well at ridiculously old age for a QB

Number of observations at each age:



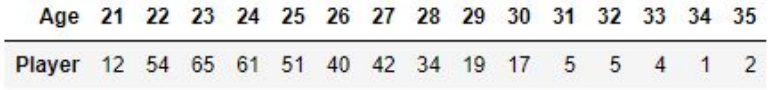
Age	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41
Player	4	7	12	11	17	15	11	11	13	12	9	10	10	8	9	8	7	4	2	1	1

EDA: RB Age



The RB graph looks great until age 31 again because of outliers. If you remove age groups with 5 or less observations it looks exactly how you'd expect.

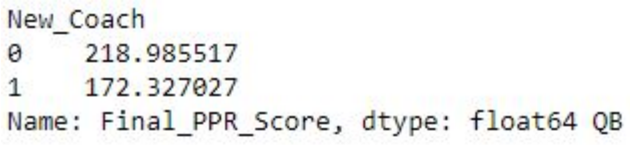
Number of observations at each age:



Age	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
Player	12	54	65	61	51	40	42	34	19	17	5	5	4	1	2

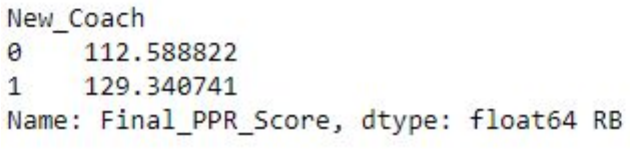
EDA: New Coach

QB:



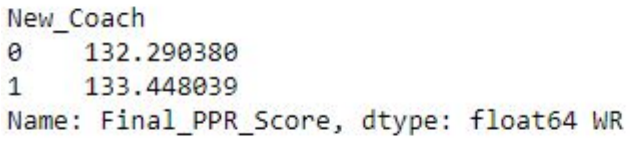
```
New_Coach
0    218.985517
1    172.327027
Name: Final_PPR_Score, dtype: float64 QB
```

RB:



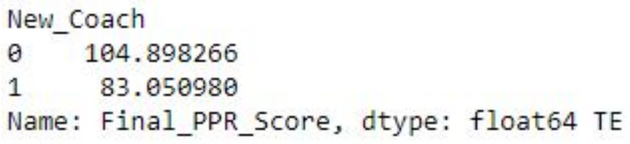
```
New_Coach
0    112.588822
1    129.340741
Name: Final_PPR_Score, dtype: float64 RB
```

WR:



```
New_Coach
0    132.290380
1    133.448039
Name: Final_PPR_Score, dtype: float64 WR
```

TE:



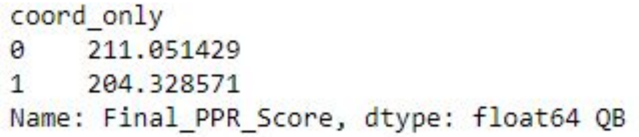
```
New_Coach
0    104.898266
1     83.050980
Name: Final_PPR_Score, dtype: float64 TE
```

Having a new coach seems to have different effects on players based on their position.

It appears to hurt QBs and TEs, help RBs, and make no difference with WRs.

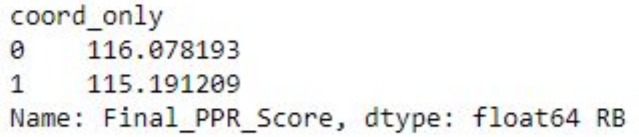
EDA: New Coordinator and NOT New Coach

QB:



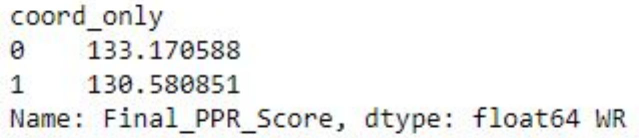
```
coord_only
0    211.051429
1    204.328571
Name: Final_PPR_Score, dtype: float64 QB
```

RB:



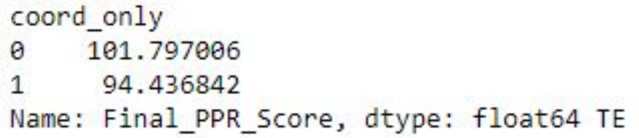
```
coord_only
0    116.078193
1    115.191209
Name: Final_PPR_Score, dtype: float64 RB
```

WR:



```
coord_only
0    133.170588
1    130.580851
Name: Final_PPR_Score, dtype: float64 WR
```

TE:

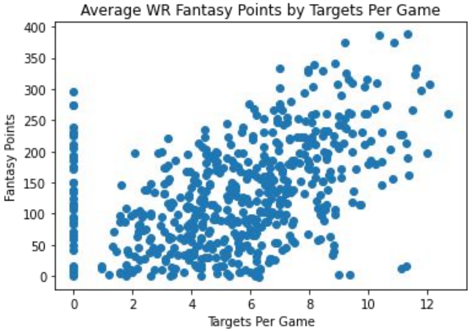


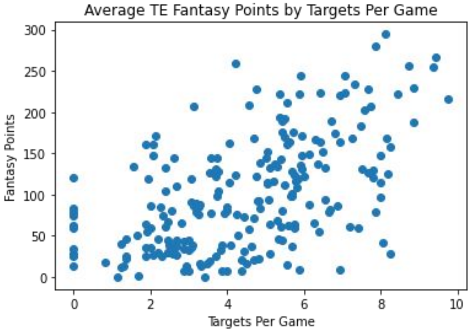
```
coord_only
0    101.797006
1     94.436842
Name: Final_PPR_Score, dtype: float64 TE
```

The trend is the same for new coordinators overall, probably related to the fact that when a team gets a new coach it almost always gets a new coordinator. However, the reverse is not true.

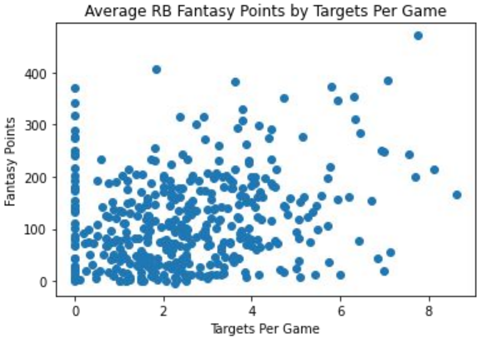
The trend is a little different for a player who only got a new coordinator. QB and TE points dip just barely, while RB and WR points stay roughly the same

EDA: Targets Per Game

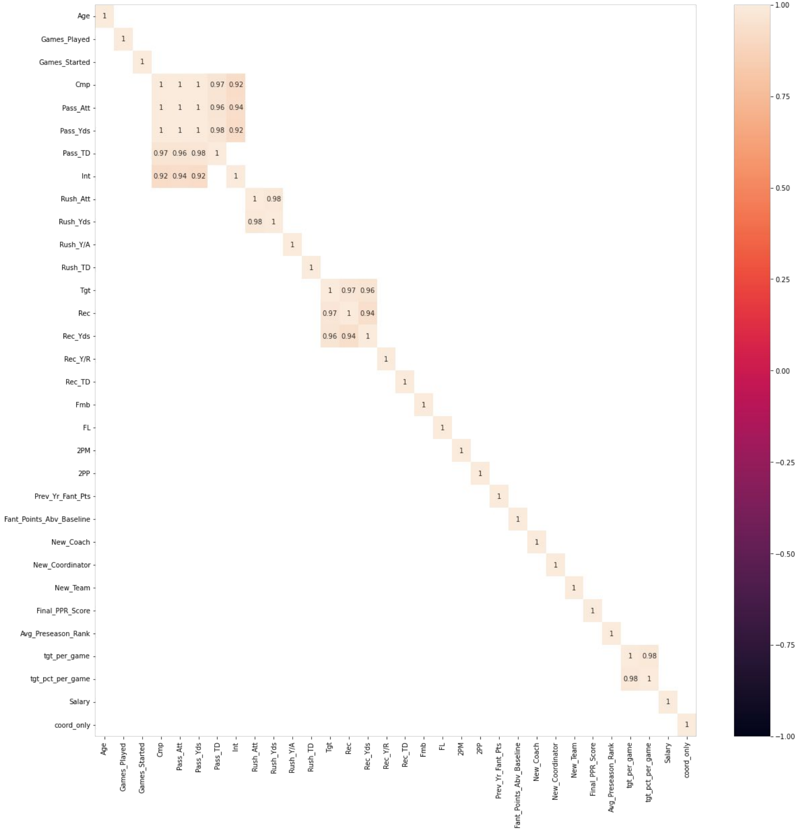




There is a pretty clear trend line for WRs and TEs targets per game. It’s not as clear for RBs but it’s still somewhat there.



EDA: Feature Correlation



Only shows correlations over 0.9

Had to remove a handful a features because of high correlations between them. Removed:

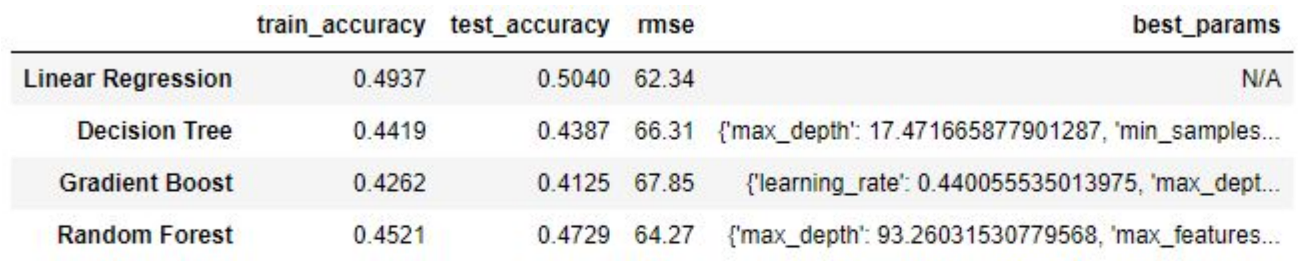
Completions, Pass Attempts, Pass TDs, Interceptions, Rush Attempts, Targets, Receptions, Targets Percentage Per Game

Modeling

Features in the model included age, passing/rushing/receiving yards, previous year fantasy points, new coach, new coordinator, new team, preseason fantasy ranking, targets per game, salary, position.

5 models total: Linear Regression, Decision Tree, Gradient Boost, Random Forest, and a Neural Network

Baseline RMSE: 88.61 (RMSE is roughly the average difference in predicted vs actual points)



	train_accuracy	test_accuracy	rmse	best_params
Linear Regression	0.4937	0.5040	62.34	N/A
Decision Tree	0.4419	0.4387	66.31	{'max_depth': 17.471665877901287, 'min_samples...
Gradient Boost	0.4262	0.4125	67.85	{'learning_rate': 0.440055535013975, 'max_dept...
Random Forest	0.4521	0.4729	64.27	{'max_depth': 93.26031530779568, 'max_features...

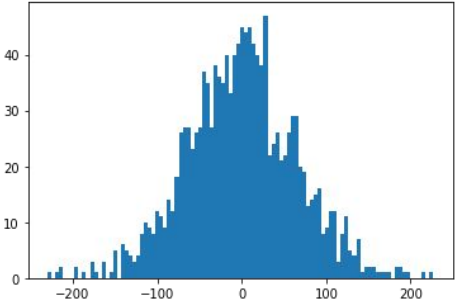
Modeling: Linear Regression

50.4% of fantasy point variance was explained by this model on data it had not yet seen

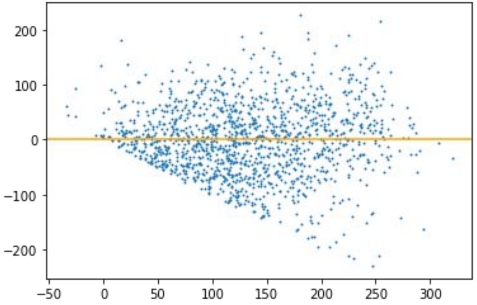
RMSE: 62.34

Assumptions Satisfied:

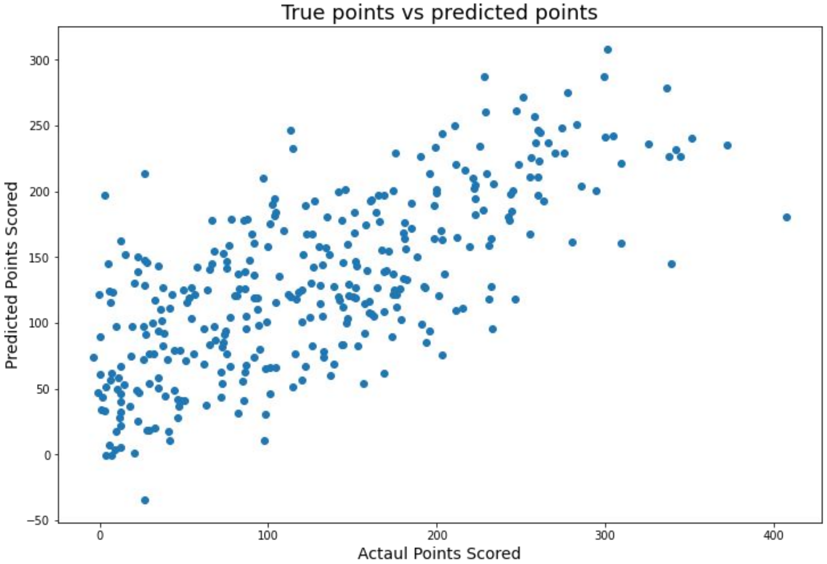
Normality of errors:



Equal Variance of Errors:



**Linearity and Independent Errors also satisfied but not pictured



Modeling: Linear Regression Coefficients

Fantasy points are expected to decrease by 2.98 for each 1 unit increase in age, holding all else constant.

For every 1 unit increase in a players preseason fantasy ranking (worse ranking) their fantasy points are expected to decrease by .398 points

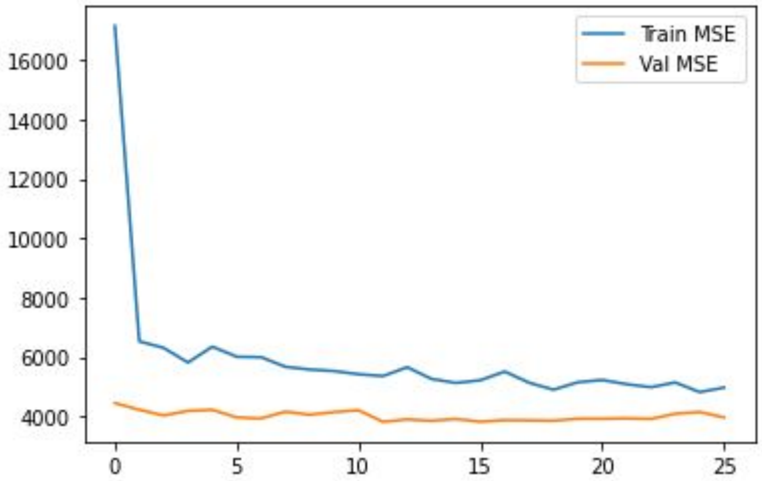
A player with a new coordinator but not a new coach, was expected to see an increase of 8.17 fantasy points, holding all else constant

Coefficients for the positions of RB, WR, and TE were -74, -68, and -71, respectively when compared to the baseline position (QB). That makes sense because QBs are the highest scoring position in fantasy football.

Modeling: Neural Network

Randomized Search CV Neural Network got the best RMSE of 61.74

Best params: 64 layer one neurons, 200 layer two neurons, 100 epochs, 0.3 dropout, and 8 batch size



Modeling: How Does it Compare to the Pros?

I found ESPNs 2019 preseason predictions (done by their fantasy sports data scientist Mike Clay) but it was a pdf file so I had to enter them into an excel sheet manually…

The RMSE for his predictions was 59.76, just barely better than mine. I gave the same advantage I had too, only predicting for the top players from that year.

Final Conclusions

It is incredibly difficult to predict season long statistics for NFL players. There is far too much variability and noise of the course of 16 NFL games (now 17) to get accurate predictions.

The current ESPN model is little bit more accurate than mine but not by much. Plus Mike Clay has all the resources of ESPN, so not exactly a fair fight. But I think that speaks to how difficult of a problem this is. Even the top professional in the industry doesn’t have a great RMSE

Improvements/Next Steps

More in depth stats - would require a paid subscription

More years worth of data - again this gets complicated when you want information other than just stats, like new coordinators and salary data

Career stats - It wasn’t possible to get a snapshot of a player’s stats in a given year for free. i.e. to get Tom Brady’s 2014 career stats to that point, I would’ve had to subtract his 2015-2021 stats from his current stats. That’s possible but it’s different for each player depending on when they quit playing. It was just too much of a lift for now.

Weekly predictions instead of season long. Less variance and noise. This is popular with professional gamblers who play daily fantasy (Draftkings and Fanduel)

Thank You!