ISyE 6740 – Spring 2024
Project Report

Team Member Names: Jackson D. Schieber (group 120)

Project Title: Data Driven Drafting

Problem Statement

In today's world, people love their sports. Some express more passion for sports than anything else in life. Thus, "fantasy" sports were invented. These fantasy sports serve as a means for people to draft their own unique team of professional or collegiate players from their chosen sport and to challenge the lineups of their friends, family, or other strangers. The game works by rewarding points to each person for the stats that players on his/her team produce. For example, if a soccer player scores a goal or gets an assist, then any people who own that player on their team will gain points. In America, one of the most popular fantasy sports is fantasy football. This game has captured immense attention from both youth and adults. One of the most popular sports TV channels – ESPN – has entire segments dedicated to fantasy football. According to USATODAY "As of last year, 29.2 million people in the United States played fantasy football, according to Statista Research. And fantasy sports in general have exploded into an $11 billion business." Therefore, people's money and pride are at stake. Unfortunately, while an objective and numerical approach to fantasy football is becoming more common, many still rely upon the subjective opinions and judgments of popular analysts. Furthermore, the numerical methods that do exist do not appear to be highly accurate. Therefore, this project aims to predict the total number of points scored by a fantasy football player in a season by accounting for several easily accessible dimensions. An accurate total points projection is incredibly helpful to people at the beginning of the season as people attempt to draft their fantasy team.

Data Sources and Preprocessing

In order to create and train the models, data is aggregated from several sources within the popular website – https://www.fantasypros.com. Additionally, pre-season rankings from ESPN analysts are also utilized and found from https://www.espn.com. Furthermore, data is gathered from the years 2016-2023. The sources from FantasyPros are structured, but the older versions of the ESPN sources are only partially so. Therefore, regex expressions are used to extract items of interest. After uploading the data and transforming the columns and values of each source to provide solely the needed information in a way that is more algorithmically digestible, the following 4 tables are produced: ESPN Rank, ADP, Points, and Stats. The first few rows of these tables can be seen below. Note that the combination of Player, Year, and Position serves as the primary key. While some players did change position, and some share the same first and last name, a negligible number share a combination of all 3 attributes. The rank table represents an average of ESPN analyst opinions on the best and worst players to have on one's team created just before players draft their teams and the season starts. This is broken down

into overall and positional rankings. Positions of interest in this analysis only include the most popular and impactful: running back RB, wide receiver WR, quarterback QB, and tight end TE. Therefore, a player of the tight end position could be ranked as 10th overall but 1st in available tight ends. The ADP table represents the average draft position of each player measured across several different popular leagues. Many of these are blank in the earlier years as some leagues were not yet originated or did not record their stats. Generally, data is more plentiful on fantasy players in more recent years. The points table represents the points of a given player in a given year. Finally the stats table houses information on a players statistical performance in the previous year including rushing attempts, rushing yards, rushing yards per attempt, longest rush, number of rushes of over 20 yards, rushing touchdowns, receptions, passing targets, receiving yards, receiving yards per reception, receiving touchdowns, fumbles, games played, longest reception, receptions over 20 yards, passing completions, passing attempts, passing completions per attempt, passing yards, passing yards per attempt, passing touchdowns, passing yards, and interceptions thrown.

Pre-season_Overall_Rank	*Pre-season_Position_Rank*	*Player*	*Year*	*Position*
1	1	David Johnson	2017	RB
81	38	Adam Thielen	2017	WR
241	74	Orleans Darkwa	2017	RB

ESPN Rank

Player	*Position*	*ESPN*	*Sleeper*	*NFL*	*RTSports*	*FFC*	*AVG*	*Year*
Antonio Brown	WR	Na	Na	Na	Na	Na	1	2016
Julio Jones	WR	Na	Na	Na	Na	Na	2.3	2016
Odell Beckham Jr.	WR	Na	Na	Na	Na	Na	2.8	2016

ADP

Player	*Position*	*AVG*	*Total*	*Year*
David Johnson	RB	25.5	407.8	2016
Aaron Rodgers	QB	23.8	380	2016
Matt Ryan	QB	21.7	347.5	2016

Points

Rushing_ATT	*Rushing_YDS*	*Rushing_Y/A*	*Rushing_LG*	*Rushing_20+*	*Rushing_TD*	*REC*	*TGT*
0	0	0	0	0	0	0	0
261	1,268	4.9	44	4	7	75	94
0	0	0	0	0	0	0	0

Receiving_YDS	*Receiving_Y/R*	*Receiving_TD*	*FL*	*G*	*Receiving_LG*	*Receiving_20+*
0	0	0	0	0	0	0

616	8.2	2	1	12	0	0
0	0	0	0	0	0	0
CMP	***ATT***	***CMP_%***	***Passing_YDS***	***Passing_Y/A***	***Passing_TD***	***INT***
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0

Player	***Position***	***Year***
Russell Wilson	QB	2017
Le'Veon Bell	RB	2017
Alvin Kamara	RB	2017

Stats

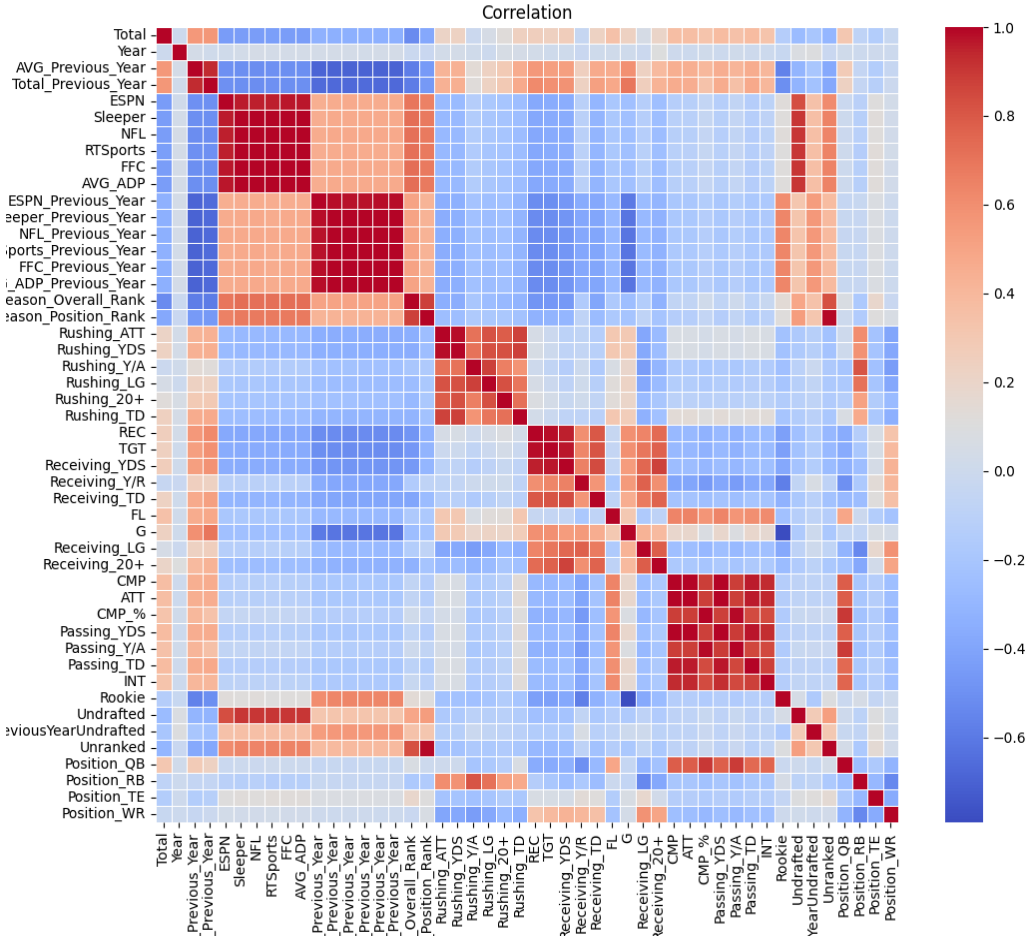
Finally, all these tables are merged into one using a left join on the points table which contains our dependent variable. ADP and the points table itself, are joined twice so as to capture last year's points scored, and last year's draft position for each player. During this process a small amount of data is lost due to discrepancies in naming convention between ESPN and FantasyPros. However, most of the inconsistencies are manually corrected. In addition to naming inconsistencies, there is some amount of missing data for the various dimensions. This is because some of the dimensions were not tracked for some of the players through FantasyPros or did not exist. Moreover, in columns that track the previous year's features for a player, there are N/A values for players in their first year. A separate indicator column called "Rookie" is created to represent this scenario. Additionally, some of the lesser-known players were not ranked, and some of them were not even drafted in the previous or present year. Columns "Unranked" and "PreviousYearUndrafted" are created to account for these conditions and the N/A values in the preseason ranking and average ADP columns are replaced with 400 and 1000. These numbers are last in each column to approximate not being picked at all. Additionally, the missing values for the website/league specific ADP columns that did not exist for certain years are replaced with the average which serves as an approximate. Finally, the position columns are converted into 4 new columns with binary representation to accommodate additional modeling capability. The features of the final product (49 columns) can be found in the appendix below along with summary statistics.

At this point it was noticed that the 25^{th} percentile of points scored on the season is just 4 points and the median is about 30. These values are substantially smaller than what relevant players output. The best player outputs 300+ points. Therefore, to eliminate noise, players with less than 40 points that were not ranked by an analyst are dropped as these players are less relevant. The new data frame is 2346 rows which is 2227 fewer than the first. All columns except for player name are transformed into either integer or float.

The predicted variable – points – has a mean, standard deviation, min, and max of 124, 85, -1, and 471 respectively (prior to train/test split). The negative number is not alarming as negative yards, fumbles, or interceptions equates to negative points. There appears to be a strong amount of variance in the dimension. The density of this feature can be observed in the KDE plots below. Additionally, there are a couple more plots exploring the different position types and the effect of year on total points. Finally, one can observe a correlation plot and some

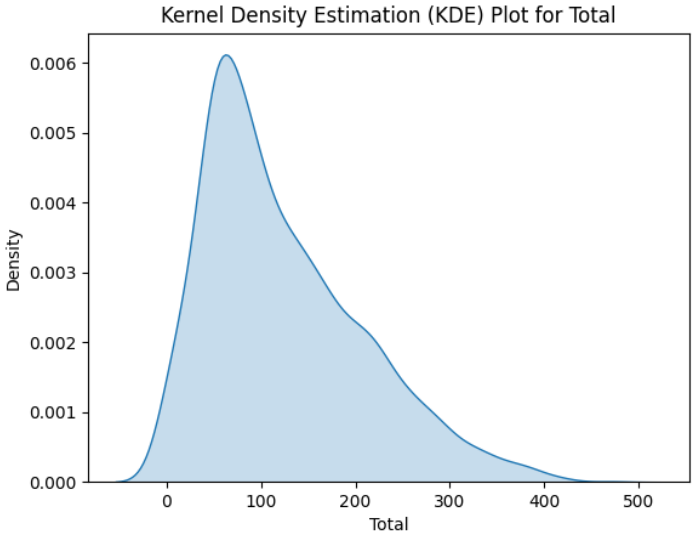
box plots of several features of interest plotted against a newly derived dimension indicating whether a player's points are more or less than the median points scored.

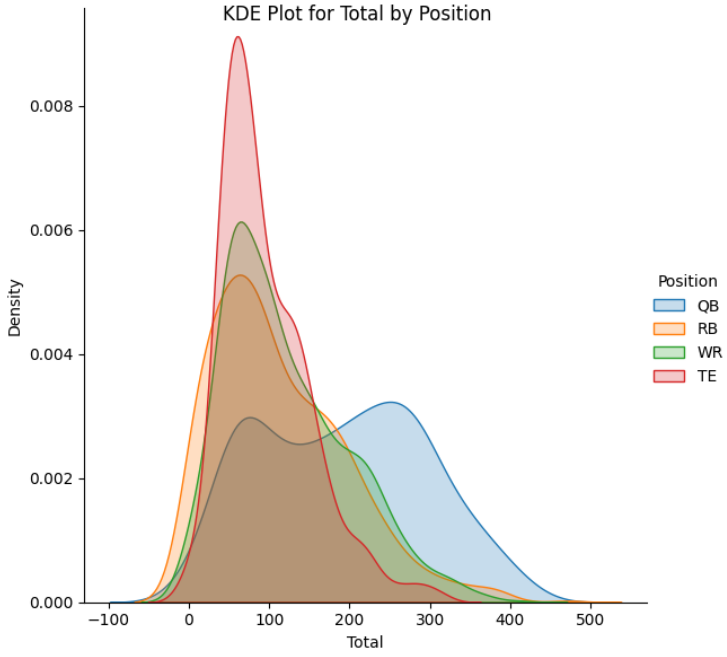
The correlation plot reveals that the different fantasy leagues have very similar ADP and probably do not all need to be represented. Looking at the rushing stats, the rushing attempts column is expectedly correlated with all other related metrics; however, it is surprising that there is a positive correlation between attempts and yards per attempt. Oftentimes, players who get less rushes are more rested and able to get an extra couple yards. Similarly, when they do get the ball, the play is sometimes specifically designed for them, and they are better positioned to excel. It appears that players who rush more are so much better than those who get fewer rushes that these effects are overruled. Receiving and passing stats are all significantly correlated so some of these attributes may not be necessary to the model. The rest of the correlations are intuitive. It is noteworthy that preseason rank and ADP are positively correlated. Also, one should remember that a desirable ADP/rank is a lower value so its negative correlation with many of the variables indicates that it is an important predictor for success. It is interesting that the total points scored in the previous year has a large positive correlation with points scored in the present year indicating that player scoring is not random and persists across time.

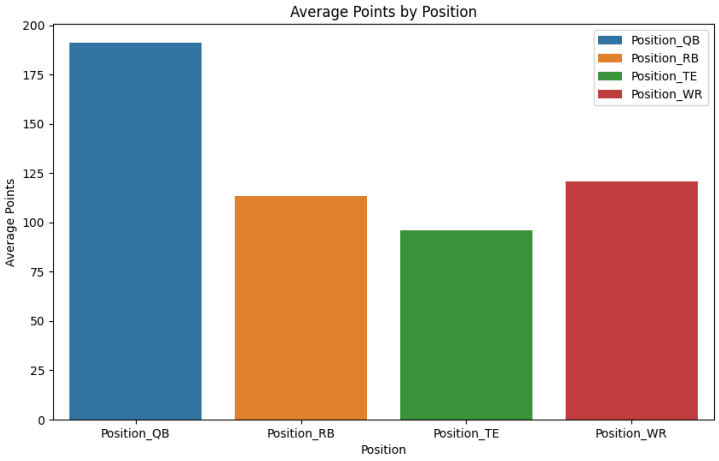


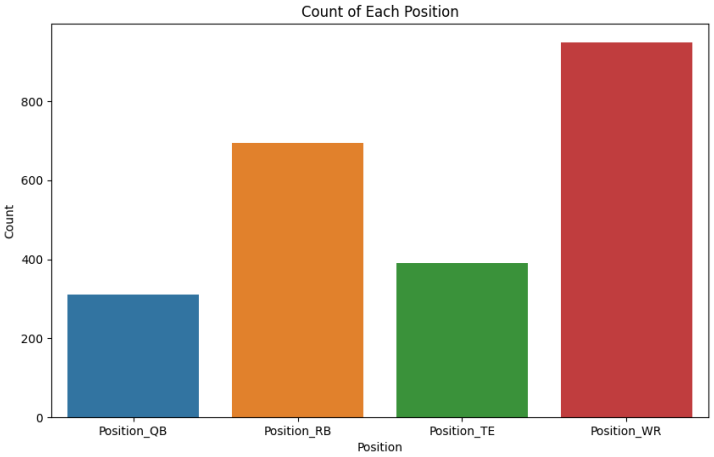
Below, overall total points demonstrates a strong right skew. Looking at the positional KDE, TE appears to have the smallest spread of points with a massive number of players scoring around 60; it has a long but thin right tail indicating that obtaining a high scoring player at this

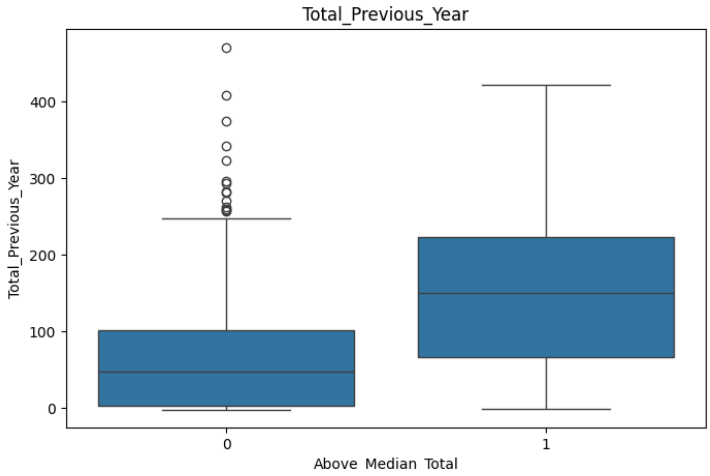
position offers high marginal value. The player is likely much better than the next player/replacement. While the shapes of RB and WR look similar, QB is the most distinct as it is more uniform. This shows how getting a top scoring player at the QB position is not as advantageous as it would be for another position. The bar plots show that WR and RB are significantly more plentiful positions than TE and QB. All of the features in the boxplots below – Total_Previous_Year, AVG_ADP, AVG_ADP_Previous_Year, Pre-season_Overall_Rank, and Pre-season_Position_Rank – have significantly different quantiles between players who scored above and below the median number. Having lower ADP/rank and scoring more points last year appear to be strong positive indicators for scoring above the median this year. However, these plots show several outliers, so these variables are not absolutely predictive. There are several players that are not even drafted in the majority of leagues (which tend to take just under 200 players in total) that obtain a total points value above the median.

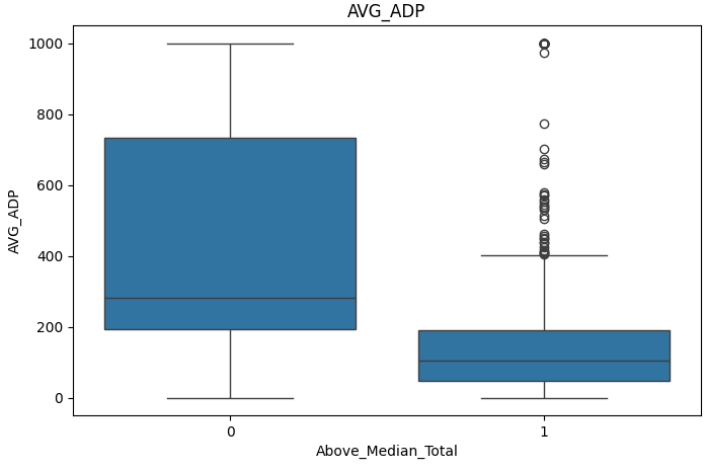


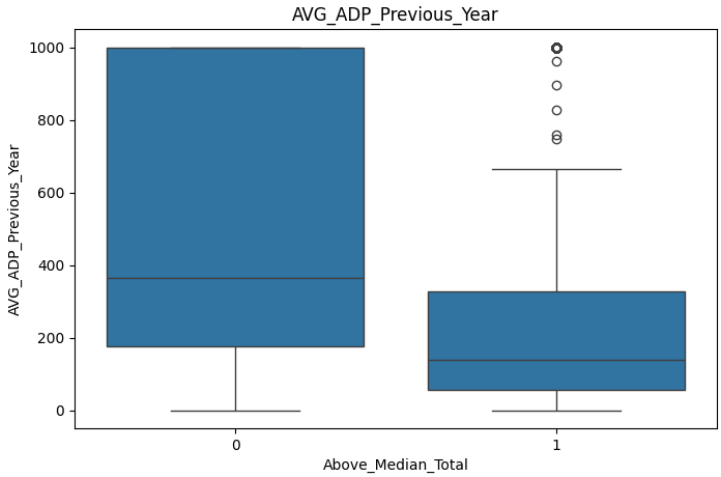


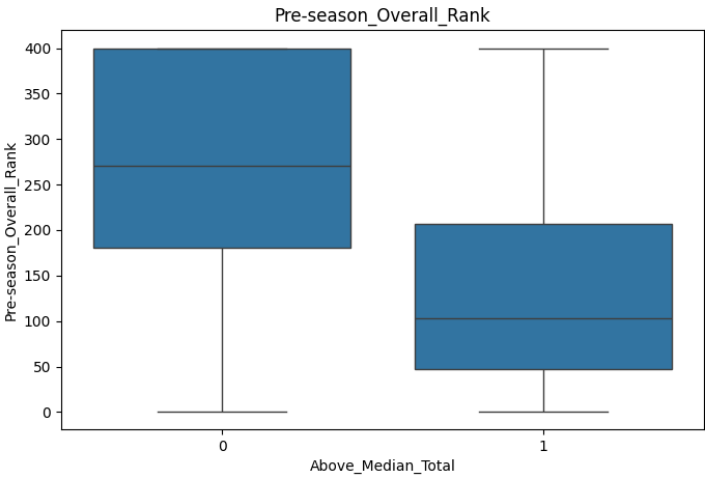

Average Points by Position
Position_QB
Position_RB
Position_TE
Position_WR
Average Points
Position

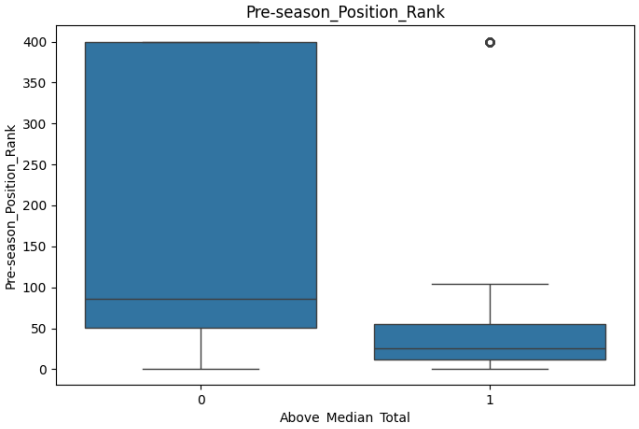

Count of Each Position
Count
Position_QB
Position_RB
Position_TE
Position_WR
Position

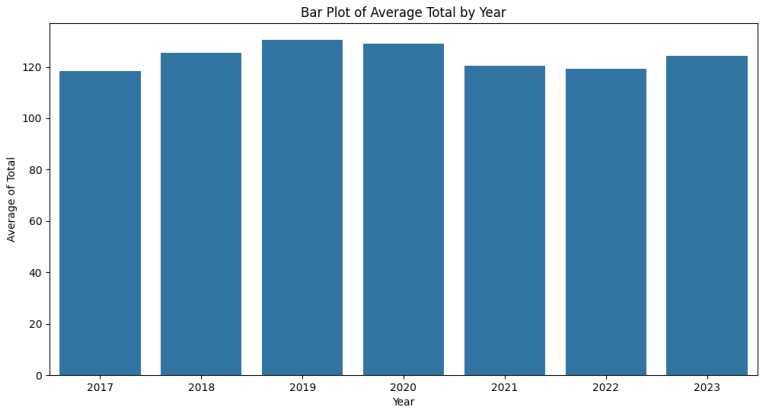

Total_Previous_Year
Above Median Total


AVG_ADP
Above_Median_Total


AVG_ADP_Previous_Year
Above_Median_Total


Pre-season_Overall_Rank
Above_Median_Total





Methodology

Prior to beginning modelling, about 15% of the data is extracted into a new "testing" data frame; In order to make sure the test set fairly represents the data, players are sequentially picked to fill the test set, from most points scored to least; the set only offers a single top scoring player (in a year), a single 2nd most scoring player, and a single 3rd most scoring player, etc.… This ensures that the test set is not heavily weighted towards high or low scoring players. Additionally, while the order/ranking is fixed, the players are randomly sampled from the different years. Therefore, after the third player is chosen the next player added to the test set must be the fourth highest scorer of his year, but the year he is chosen from is randomized. This is preferred to simply taking a single year of player scoring data because one year's scoring can be biased. For example, some years are simply higher scoring than others as seen in the bar plot above. Also, for the purposes of modeling, the player name and year columns are dropped.

Several models including linear regression, support vector machine (kernels = poly/rbf/sigmoid), random forest, gradient boosting, K-nearest-neighbor, lasso/ridge regression, and XGBoost are utilized. Had it not taken so long to run, a linear kernel also would have been incorporated with support vector machine (SVM). Mean squared error (MSE), mean absolute deviation (MAD), and R-squared are measured and recorded for each. The best and most desirable models will produce a combination of the highest R-squared and lowest MSE/MAD. Additionally, the hyperopt package is used to assist in hyperparameter tuning of the XGBoost regression. This process is repeated for both unscaled and scaled data. It is also repeated a third time for a subset of the original data where several of the columns with high correlation are removed.

Evaluation and Final Results

In the table below, one can see the recorded metrics of each model produced from the non-scaled and scaled full dataset. Note that MSE_Top_200 represents the MSE of just the top 200 highest scorers in the data which might be more representative of the kind of players someone is more likely to draft before the season starts. Finally, a third table is provided with the results of the model after removing high correlation features. SVR performs quite poorly with all kernels though the linear kernel may have had more success since linear regression was a top performer. The average MSE for all scenarios ranges between 4100 and 4200. Therefore, scaling beforehand (some techniques already scale) and reducing dimensionality by removing

high correlation features does not provide significant improvements. This may be evidence that the models are not overfitting. KNN is the next worst performer, followed by XGBoost. Linear regression, ridge and lasso regression, random forest, and gradient boosting all perform the best with R-squared values between 46 and 48 and MAD's of about 45. Considering the wide dispersion of points scored ranging from about 0 to 470 and the STD of 85, the MAD value feels successful. After optimizing hyperparameters, XGBoost produces a total MSE of 4100 and an MSE of 4497 among just the top 200 highest scorers. This is for the full set of data. Some optimized parameters include: colsample_bytree - 0.9026952470199305, gamma - 5.073901819670221, max_depth - 4.0, min_child_weight - 2.0, reg_alpha - 165.0, reg_lambda - 0.0423565250139723. This score is actually worse than what is found in the third table below, but that is likely because that third table is modeled with a subset of the total features. Changing these parameters does improve upon the full feature model slightly, but the difference is small. The mediocre performance is strange considering that XGBoost is short for extreme gradient boosting, and the gradient boosting regressor performed well. This suggests that these hyperparameters may not have been "optimized" effectively. More care is warranted for parameter optimization in the future.

Observing the coefficients found in the table below of some of the highest scoring models – linear regression and ridge regression – one can get a feel for how the features effect total points. For example, the binary feature indicating position type awards the most points for QB and actually penalize TE and RB. This is consistent with the average points scored of the positions. Similarly rushing/receiving yards and receptions increase the total predicted points scored. Interestingly rushing attempts and passing targets are strongly negative suggesting that the model is penalizing low efficiency. Strangely, this is reversed with the QB position as the model penalizes completions but rewards passing attempts. This could be because QB's who are more aggressive with their passes will gain more yards and touchdowns on average but will throw more incomplete passes. Some quarterbacks may have many completions, but they might be safe and short. Passing yards is strongly positive, but passing touchdowns is actually penalized. While this penalization of touchdowns is very counterintuitive, it makes sense in the context of the reward given for passing yards. A quarterback is not going to score touchdowns without also obtaining yards. Furthermore, touchdowns are dependent upon randomness and luck with a higher variance than yards thrown. Therefore, having many touchdowns the previous season may actually overinflate consensus opinion on that player's performance this year since the player is likely to regress back to the mean. There are a few other strange coefficients that can probably be explained with similar logic by looking at the context of the other features. For example, fumbles (rewards negative points) has a positive coefficient. Other variables like Sleeper and FFC (ADP from two different leagues) have very strong positive and negative coefficients, but they represent very similar things. This is likely a product of having too many correlated variables; however, average ADP does have a strong negative correlation which lends credibility to the results. Overall, ridge and standard linear regression assigned similar coefficients to the features. There are very few that ridge reduced near zero that linear regression did not also have as very small. The biggest difference between the two is that pre-season position rank is strongly positive for ridge regression and negative for the standard approach.

Model	*MSE*	*MAD*	*R-squared*	*MSE_Top_200*
LinearRegression	3614.633	45.085595	0.474189	3933.244676
RandomForestRegressor	3591.987	45.22278	0.477484	3921.667434
GradientBoostingRegressor	3691.228	45.068557	0.463047	4081.656465
SVR(kernel='poly')	5855.363	56.098374	0.148237	7356.230605
SVR(kernel='rbf')	4585.376	51.005165	0.332978	5719.562273
SVR(kernel='sigmoid')	5314.819	55.517768	0.226868	6359.925558
XGBRegressor	4185.875	48.252939	0.391092	4575.580429
KNN	4387.214	50.714102	0.361804	4919.437667
Lasso	3801.259	46.153628	0.447041	4083.533427
Ridge	3609.959	45.033377	0.474869	3924.949557

Results Unscaled

Model	*MSE*	*MAD*	*R-squared*	*MSE_Top_200*
LinearRegression	3614.633	45.0856	0.474189	3933.245
RandomForestRegressor	3576.594	45.19691	0.479723	3923.995
GradientBoostingRegressor	3694.607	45.10797	0.462556	4079.11
SVR(kernel='poly')	5073.169	51.80856	0.26202	6205.969
SVR(kernel='rbf')	4688.418	51.2588	0.317989	5974.842
SVR(kernel='sigmoid')	4554.785	50.80775	0.337428	5474.686
XGBRegressor	4185.875	48.25294	0.391092	4575.58
KNN	4358.975	49.15336	0.365912	4861.025
Lasso	3666.496	45.83394	0.466645	3885.896
Ridge	3614.769	45.11531	0.47417	3891.252

Results Scaled

Model	*MSE*	*MAD*	*R-squared*	*MSE_Top_200*
LinearRegression	3671.178	45.19083	0.465964	3931.452
RandomForestRegressor	3756.114	46.23872	0.453609	4037.744
GradientBoostingRegressor	3660.131	45.03087	0.467571	3964.023
SVR(kernel='poly')	5161.024	51.69182	0.24924	6326.634
SVR(kernel='rbf')	4689.055	51.24813	0.317896	6007.215
SVR(kernel='sigmoid')	4393.785	49.75036	0.360848	5189.056
XGBRegressor	4014.71	47.9677	0.415991	4570.667
KNN	4418.995	48.94285	0.357181	4914.738
Lasso	3669.696	45.7124	0.46618	3877.947
Ridge	3667.083	45.09568	0.46656	3920.766

Results Fewer Features

Variable	***Coefficient LR***	***Coefficient Ridge***
AVG_Previous_Year	1.209225	0.192037
Total_Previous_Year	7.240371	16.252487
ESPN	0.835077	-8.695324
Sleeper	236.187	146.475815
NFL	39.07671	20.751894
RTSports	-2.12754	-25.055336
FFC	-148.292	-77.530123
AVG_ADP	-148.292	-77.530123
ESPN_Previous_Year	-2.26283	1.70236
Sleeper_Previous_Year	-38.1044	6.48962
NFL_Previous_Year	-7.46976	0.638
RTSports_Previous_Year	-38.7664	-22.85507
FFC_Previous_Year	46.72326	10.333792
AVG_ADP_Previous_Year	46.72326	10.333792
Pre-season_Overall_Rank	-29.4961	-32.756466
Pre-season_Position_Rank	-115.952	-100.311687
Rushing_ATT	-19.5201	-19.456616
Rushing_YDS	22.63834	20.176381
Rushing_Y/A	-1.03845	-0.778419
Rushing_LG	-1.28711	-1.454976
Rushing_20+	-2.93395	-2.878379
Rushing_TD	-0.50933	-1.764604
REC	9.072345	6.648468
TGT	-28.0492	-26.986394
Receiving_YDS	17.71018	14.366282
Receiving_Y/R	-4.31844	-4.36342
Receiving_TD	2.516616	1.504022
FL	2.461134	2.733641
G	7.336342	6.666483
Receiving_LG	-0.099	-0.03876
Receiving_20+	-1.23714	-1.206584
CMP	-80.9848	-55.181871
ATT	26.19201	13.429908
CMP_%	-19.3825	-23.483783
Passing_YDS	65.89648	47.738168
Passing_Y/A	18.62974	23.200942
Passing_TD	-3.92056	-5.587547
INT	-3.22163	-1.708176

Rookie	-0.72021	-0.783846
Undrafted	13.22226	11.840771
PreviousYearUndrafted	-3.25538	-2.910359
Unranked	123.9437	110.952931
Position_QB	12.74525	13.511405
Position_RB	-5.63218	-5.798974
Position_TE	-8.65944	-8.091316
Position_WR	2.800659	2.000423

Model Coefficients

As mentioned above, the R-squared is not as high as hoped. This is partly because football performance is simply random. High caliber players get injured and miss entire seasons. Consequently, unknown players step into the missing player's spot and heavily outperform expectations. Injuries are likely the biggest cause of randomness. Even if a player stays healthy, if his QB gets hurt, his value can be ruined. Additionally, players get transferred to new teams mid-season or refuse to play as a form of contractual negotiation. No set of lineups or player is ever the same from season to season and sometimes the emergence of a new player strongly elevates or reduces the point output of another player on that same team. Therefore, it's difficult to predict future performances. However, I still think that the models can meaningfully improve upon these results. In the future, it would be beneficial to simply acquire more data. This analysis tracked points from 2016-2023 but this is likely not sufficient to fully understand the data given its complexity and randomness. It may be easy for a model to mistake a random event for a pattern without enough of a sample. Additionally, the parameters in these models could likely be improved. Only parameters in XGBoost were adjusted, and its underperformance to the generic gradiant boosting algorithm indicates misassigned parameter values. An additional future effort that would be beneficial includes creating a scorecard that evaluates the predictions relative to the analyst predictions. This is a bit difficult with the way the data set is constructed since there can be several players ranked number 1 in the testing dataset because it is constructed from all of the year's data. Though these models have their flaws, they appear robust enough to offer real guidance to players and analysts and would likely outperform people who pick based on gut feeling or those who are new to fantasy football. Moreover, future improvement efforts could yield models that outperform most other predictive methods.

Appendix

Gardner, S. (2023, December 20). *Money. power. women. the driving forces behind Fantasy Football's skyrocketing popularity.* USA Today. https://www.usatoday.com/story/sports/nfl/fantasy/2023/12/15/fantasy-football-sports-economy/71870731007/

	mean	***std***	***min***	***25%***	***50%***	***75%***	***max***
Total	124	85	-1	58	102	174	471
Year	2020	2	2017	2018	2020	2022	2023
AVG_Previous_Year	8	6	0	2	8	13	30
Total_Previous_Year	107	96	-2	17	90	171	471
ESPN	313	329	1	89	193	336	1000
Sleeper	297	311	1	86	192	328	1000
NFL	299	310	1	90	199	315	1000
RTSports	294	312	1	85	189	317	1000
FFC	298	311	1	87	192	329	1000
AVG_ADP	298	311	1	87	192	329	1000
ESPN_Previous_Year	422	397	1	95	226	1000	1000
Sleeper_Previous_Year	410	388	1	93	229	1000	1000
NFL_Previous_Year	411	387	1	99	235	1000	1000
RTSports_Previous_Year	408	389	1	91	227	1000	1000
FFC_Previous_Year	410	388	1	93	228	1000	1000
AVG_ADP_Previous_Year	410	388	1	93	228	1000	1000
Pre-season_Overall_Rank	205	135	1	86	197	295	400
Pre-season_Position_Rank	129	157	1	22	53	103	400
Rushing_ATT	33	65	0	0	1	33	378
Rushing_YDS	147	292	-14	0	3	125	2027
Rushing_Y/A	1	2	0	0	0	0	11
Rushing_LG	8	18	0	0	0	0	99
Rushing_20+	1	3	0	0	0	0	37
Rushing_TD	1	3	0	0	0	1	18
REC	27	29	0	0	18	44	149
TGT	40	43	0	0	27	66	191
Receiving_YDS	304	358	-3	0	175	490	1947
Receiving_Y/R	8	6	-3	0	9	12	41
Receiving_TD	2	3	0	0	1	3	18
FL	1	1	0	0	0	1	11
G	11	6	0	6	14	16	17

Receiving_LG	21	26	0	0	0	42	98
Receiving_20+	5	9	0	0	0	8	57
CMP	27	89	0	0	0	0	485
ATT	42	138	0	0	0	0	719
CMP_%	7	20	0	0	0	0	91
Passing_YDS	309	1011	0	0	0	0	5316
Passing_Y/A	1	2	0	0	0	0	10
Passing_TD	2	7	0	0	0	0	50
INT	1	3	0	0	0	0	21
Rookie	0	0	0	0	0	0	1
Undrafted	0	0	0	0	0	0	1
PreviousYearUndrafted	0	0	0	0	0	0	1
Unranked	0	0	0	0	0	0	1
Position_QB	0	0	0	0	0	0	1
Position_RB	0	0	0	0	0	1	1
Position_TE	0	0	0	0	0	0	1
Position_WR	0	0	0	0	0	1	1

Summary Statistics of Data