THE FANTASY DEFENSE

Seth Galluzzi, Maxwell Jones, Said Mrad
DS 6050
Group 3
University of Virginia
Charlottesville, VA

ABSTRACT

In this exploration, we aimed to predict a team's defense/special team's fantasy performance using statistics from their previous season. To do so, we used data from FantasyDataPros containing regular season NFL data from 2002-2019 as well as some feature engineered variables to train several machine learning models such as linear regression, random forest, and gradient boosting models. Our best model, a linear regression model, with a regular parameter of 0.25, a net parameter of 0.8, and 10 iterations, yielded an MSE of 14098 and correctly predicted 4 of the top 10 defense/special team (D/ST) units in the 2019 NFL season. Given that the model outperformed several competitor predictions, we deem our model to be usable and our objective a success. Furthermore, by comparing the predictions of all the models we identified the Buffalo Bills, Minnesota Vikings, and Baltimore Ravens as our top defensive choices for the 2019 season. According to FantasyPros.com, all three teams were top ten defenses in 2019.

1 Data and Methods

1.1 Introduction

Fantasy sports, a gaming industry where participants select real-life players and are rewarded with points based on their performances, is a growing industry, expected to reach 44.07 billion dollars by 2027 according to Yahoo Entertainment [1], largely coming from the growing popularity of players, growing investments in digital and internet infrastructures, and the rise of fantasy sports applications. Fantasy football especially has become extraordinarily popularly–Fox Sports [2] mentions that the fantasy NFL industry alone would be worth 70 billion dollars right now, as Forbes reports that about 32 million Americans will spend around 15 billion dollars just on league fees, website prize fees, and information materials, more money than the NFL makes through TV rights as well as ticket and merchandise sales. This of course does not even account for indirect expenditures such as time spent by players setting up lineups, checking performances of their team, and doing research on trades and waivers.

Given the popularity of fantasy football, many people care a great deal about which players to draft, as better players will yield them more points. In most leagues, each team consists of seven offensive players, a kicker, and a defense/special teams (D/ST). As offensive players generate the majority of a team's points, people spend most, if not all, of their time deciding which offensive players to draft. While there are still many surprises each year in terms of the best and worst offensive players, most people have a general idea of which players to prioritize and which ones to avoid. However, D/ST is a less-researched area. Thus, gaining a better understanding of D/ST could prove advantageous on draft day.

The aim of this exploration is to gain a deeper understanding of D/ST, and to determine if previous D/ST data can be used to predict future D/ST results. We also hope that our research will prove useful in the pursuit of fantasy football glory. D/ST is an incredibly volatile position. The score is measured by how well a team performs on defense and special teams in the NFL. Less points and yards allowed by a defense translates into more fantasy points. Big plays such as turnovers, sacks, and touchdowns gain extra points for the D/ST. To accomplish our aim, we used each NFL team's previous season's defensive data to predict how their next season's defense would perform in terms of fantasy football. We built machine learning models such as linear regression, random forest, and gradient boosting regression to make predictions. Finally, we compared the model MSEs, analyzed visualizations, and determined the best model to predict future D/ST success.

1.2 Data Source and Variables

To train our models, we sampled fantasy D/ST data from the 2002-2019 regular seasons, obtained from FantasyDataPros at ***https://github.com/fantasydatapros/data***. FantasyDataPros provides weekly fantasy stats dating back to 1999 and yearly fantasy stats dating back to 1970. For our project, we decided to only use yearly data from 2002-2019 for several reasons, which we will discuss in the data cleaning section. In Figure 1, we ordered the data by team and year to create our response variables based on each team's following season.

	Tm	PointsAllowed	TotalYardsAllowed	OffensivePlaysAllowed	YardsPerPlay	TO	ForcedFumbles	TotalFirstDownsAllowed	Cmp	PassingAttAllowed	...
28	Arizona Cardinals	417.0	6020.0	1046.0	5.8	25.0	8.0	335.0	335.0	535.0	...
31	Arizona Cardinals	452.0	5504.0	993.0	5.5	23.0	10.0	326.0	311.0	497.0	...
11	Arizona Cardinals	322.0	5141.0	993.0	5.2	30.0	15.0	282.0	271.0	505.0	...
25	Arizona Cardinals	387.0	4729.0	936.0	5.1	26.0	11.0	272.0	301.0	488.0	...
28	Arizona Cardinals	389.0	5591.0	1018.0	5.5	33.0	17.0	331.0	321.0	522.0	...

Figure 1: Sample of First Five Rows of dataset

For our response variable, we created new **Next_PPR_Allowed** and **Next_PPR_Rank** variables, each one representing an NFL team's D/ST points per reception (PPR) fantasy performance in the following season, with Next_PPR_Allowed representing the D/ST numerical points allowed and Next_PPR_Rank representing the D/ST rank (1-32). For our predictor variables, while we had many D/ST statistics available, after early data analysis with our response variables, we used only those listed in Figure 2:

Variable	**Response/Predictor**	**Description**
Next_PPR_Allowed	Response	PPR points allowed in the following season
Next_PPR_Rank	Response	PPR rank (1-32) in the following season

PointsAllowed	Predictor	Actual points allowed
TotalYardsAllowed	Predictor	Yards allowed by the defense
OffensivePlaysAllowed	Predictor	Offensive plays allowed
YardsPerPlay	Predictor	Average yards per offensive play allowed by the defense
TO	Predictor	Turnovers forced by the defense
ForcedFumbles	Predictor	Fumbles forced by the defense
TotalFirstDownsAllowed	Predictor	Total first downs allowed by the defense
Cmp	Predictor	Total passing completions allowed by the defense
PassingAttAllowed	Predictor	Passing attempts allowed by the defense
PassingTDAllowed	Predictor	Passing touchdowns allowed by the defense
PPRFantasyPointsAllowed	Predictor	Points per reception fantasy points allowed by the defense

Figure 2: Response and Predictor Variables

1.3 Data Cleaning

To make the data more usable for modeling, we had to look at a number of changes in NFL data over its long history. First, the NFL has had many changes to its number of games; originally set at 14 in 1960, the games increased to 16 in 1978, a strike season shortened the 1982 season to 9 games, and the final change increased the games from 16 to 17 in 2021. Additionally, the number of teams has increased to 32, with the Houston Texans joining to become the 32nd and most recent team.

Aside from the number of games and team changes, in-game trends have changed drastically as well. Due to rule changes and play calling changes, offenses have become much more pass-centric. As we can see in Figures 3 and 4, the teams before 1980 tended to run the ball more and more each season, while the run game since 1980 has become quite stagnant. On the other hand, while teams before 1980 seemed to be passing less and less each season, teams since 1980 have put much more focus on passing the ball.

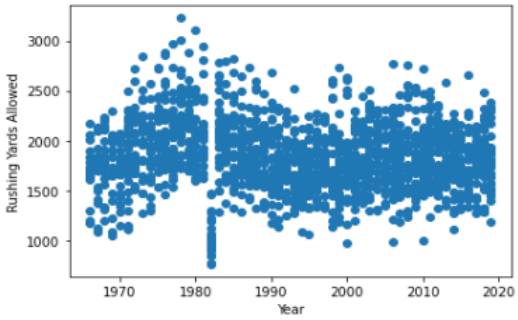


Figure 3: Rushing Yards Allowed

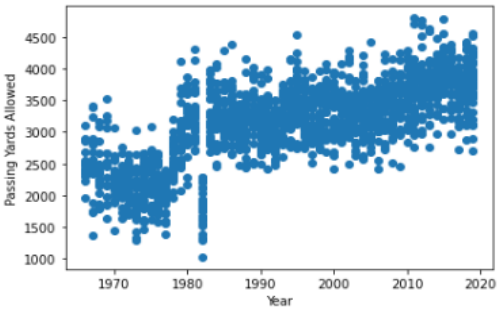


Figure 4: Passing Yards Allowed

Given the inconsistencies in the data over time, we decided to use the data from 2002 to 2017 for training and testing purposes. We held out the 2018 data to use as a validation set. Finally, we discarded the 2019 data because the response variables were incorrect. To note, the 2019 data was identified as incorrect during the modeling phase, and discarding it drastically improved the MSEs of our models. By using data from 2002 to 2017, we did not have to worry about the different number of games per season, team name changes, or outdated trends of poor passing and great running while making our predictions. Thus, after creating the predictor variables Next_PPR_Allowed and Next_PPRDefRank and restricting our data from 2002 to 2017, we were ready to begin modeling.

2 Modeling

Before exploring our models, we first gathered our feature variables using VectorAssembler and scaled them using StandardScaler. The data was split 80-20 for training and testing purposes. Many different models were built and tinkered with during the modeling process. We chose to highlight the four models below because they had some of the best results, and they help illustrate the many different types of models we tried. The first three models are regression models that use Next_PPR_Allowed as a response variable. The last model is a classification model that uses Next_PPRDefRank as a response variable. The models will be compared by using the mean squared error of the testing data, analyzing visualizations of the predictions, and exploring predictions of the 2019 season through the use of our 2018 validation set.

2.1 Linear Regression Model

The first regression model highlighted is a linear regression model used to predict Next_PPR_Allowed. The model had 10 iterations, a regular parameter value of 0.3 and an elastic net parameter of 0.8. The model had an MSE of 14099. Figure 5 illustrates the predictions versus actual Next_PPR_Allowed values of each team within the testing set.

The model appears to be a bit conservative and maintains predictions between 1100 - 1500 points allowed. Furthermore, most predictions fall between 1200 - 1300 points allowed. With that being said, the model still does a decent job making predictions.

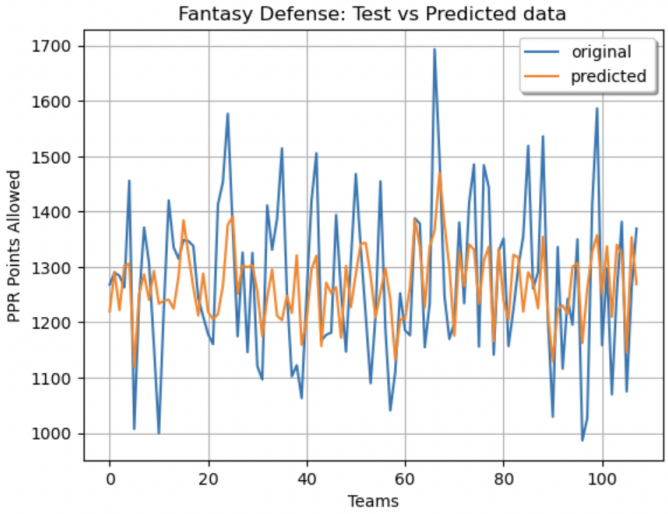


Figure 5: Linear Regression Predictions

2.2 Random Forest Regression

The second model highlighted is a random forest regression model used to predict Next_PPR_Allowed. The model had an MSE of 16215. Figure 6 illustrates the predictions on the testing data appears to be similar to the linear regression. However, it is interesting to note that the random forest regression model did not predict any values greater than 1400 points, and seemed more inclined to predict lower values than the linear regression.

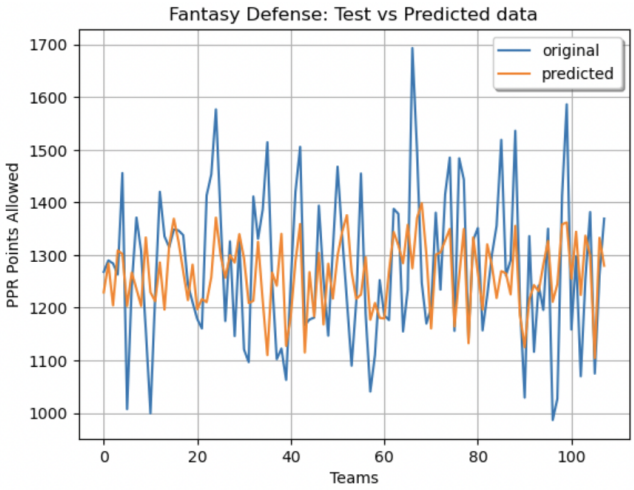


Figure 6: Random Forest Regression Predictions

To better understand these predictions, Figure 7 depicts the first 5 predictions of our testing set. To note, some of the predictions, like the 2nd entry, are extremely close to the

actual points allowed. However, it is also important to note that the predictions are all in the 1200s which is not surprising considering Figure 6 shows most values falling within the 1200-1300 range.

Tm	Next_PPR_Allowed	prediction
Arizona Cardinals	1267.86	1227.2988718923286
Arizona Cardinals	1289.86	1287.8693049542494
Arizona Cardinals	1284.02	1225.8374703935403
Atlanta Falcons	1262.9	1296.3742768946124
Atlanta Falcons	1455.88	1285.6378042819956

Figure 7: First Five Random Forest Predictions

2.3 Gradient Boosting Regression

The third model we chose to highlight was the gradient boosting regression model. The model had an MSE of 20792. Based on the MSE, the gradient boosting regression model appeared to perform the worst. However, it is interesting that, although the model had a higher MSE, the predictions from the model seemed to be a bit more bold than the previous models' predictions. The model predicted a few values to be below 1100 points, and even had a prediction near 1500 points. Fortune favors the bold in fantasy football and this model is tempting, but we believe the improved MSEs of the other two models might make them more preferable.

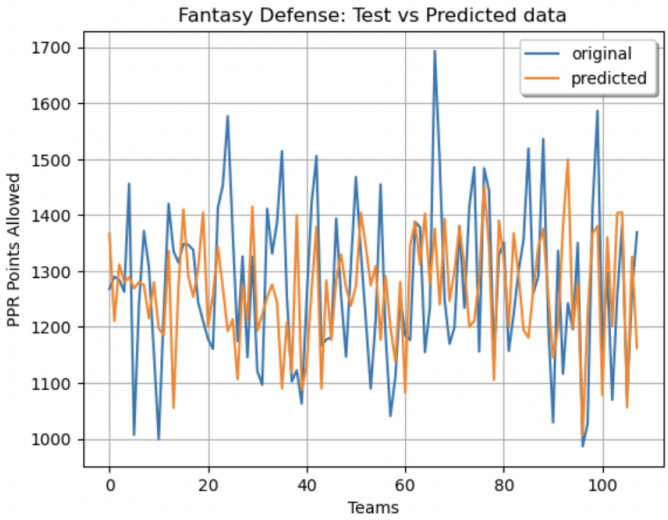


Figure 8: Gradient Boosting Regression Predictions

2.4 Random Forest Classifier

The last model we chose to highlight is a little different than the previous three. This model is a random forest classifier that predicts the Next_PPRDefRank. The model's MSE of 116 should not be compared to the previous models. It can be expected that a classifier with 32 different options is going to have its flaws. Some of these flaws

can be seen in Figure 9 and Figure 10. Another issue with this model is that the model could potentially predict teams of the same year to have the same ranking. With all that being said, the classifier did an ok job predicting the Next_PPRDefRank.

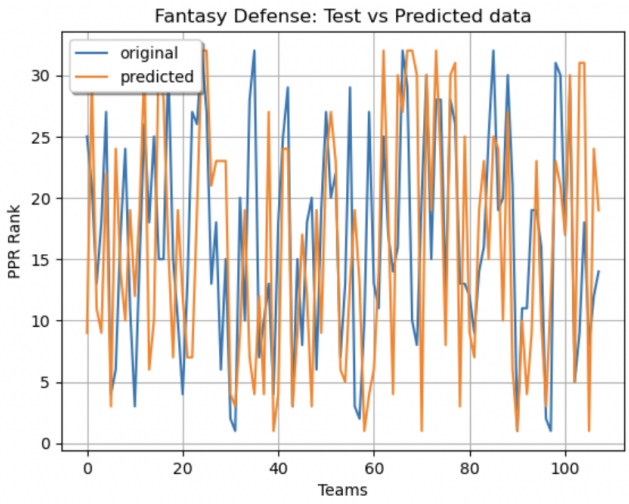


Figure 9: Random Forest Classifier Predictions

Tm	Next_PPRDefRank	prediction
Arizona Cardinals	25.0	17.0
Arizona Cardinals	21.0	23.0
Arizona Cardinals	13.0	11.0
Atlanta Falcons	18.0	23.0
Atlanta Falcons	27.0	15.0

Figure 10: First Five Random Forest Classifier Predictions

3 Results

3.1 Model Tuning and Performance

The models above were tuned in an effort to improve model performance in a variety of ways. First, a parameter grid and cross validation were incorporated to improve the performance of the linear regression model. Using ParamGridBuilder, many different combinations of parameters, hyperparameters, and iterations were explored. CrossValidator was used to create 5-fold and 10-fold models, and it was determined that the 5-fold model was superior to the 10-fold model in both performance and runtime. Unfortunately, the use of a parameter grid and cross validation did not improve the model. In fact, the mean squared error of the linear regression model using the parameter grid and cross validation increased to 14212. Figure 11 illustrates the predictions on the testing data of this model appears to be similar to the original linear regression model.

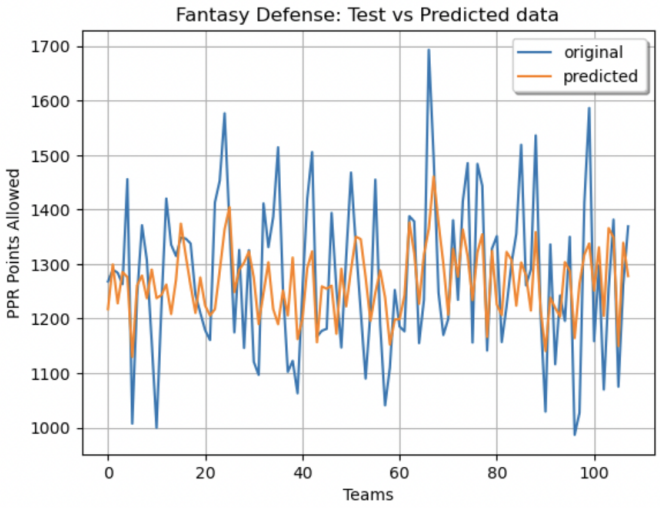


Figure 11: Linear Regression Predictions After Tuning

While exploring the parameter grid and cross validation, it was deduced that a small improvement in performance could be achieved by changing the regular parameter of the linear regression model from 0.3 to 0.25. However, the miniscule improvement in mean squared error from 14099 to 14098 felt more like a moral victory, rather than a superb improvement in results.

The random forest classification model predicting Next_PPRDefRank was also modified to improve model performance. One of the major issues of the random forest classifier model was that there were 32 options for rankings. To simplify the model, the variable PPRDefRankSplits was engineered to reduce the number of options to 4. Thus, if a team's rank was from 1 to 8 the PPRDefRankSplits value was a 1, if a team's rank was from 9-16 the PPRDefRankSplits value was 2, if a team's rank was from 17 to 24 the PPRDefRankSplits value was 3, and if a team's rank was from 25 to 32 the PPRDefRankSplits value was 4. In a sense, the new model predicts which quartile the team will rank rather than the specific rank. Figure 12 illustrates the model's predictions.

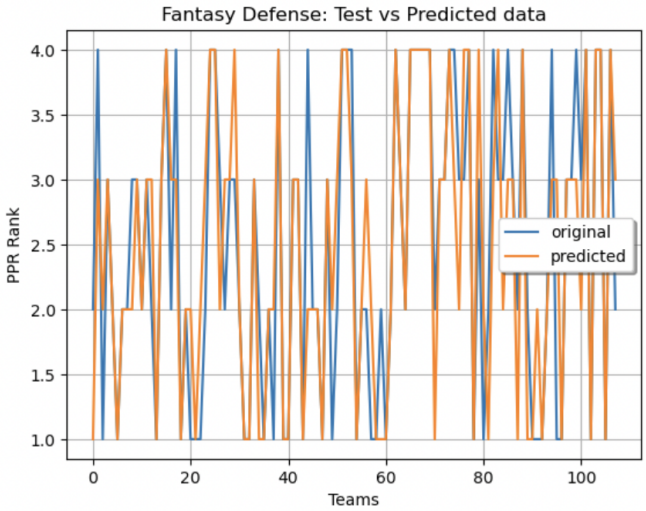


Figure 12: Random Forest Classifier Predictions After Tuning

The model had an MSE of 273. However, this model could be considered an improvement because within the constructs of fantasy football, choosing a top 8 defense from 32 teams is more realistic than choosing the top defense out of 32. Thus, by focusing on teams that received values of 1 for PPRDefRankSplits, we can narrow our search when identifying which defense to choose. The idea that narrowing our search of teams was more realistic than identifying the number one defense was also incorporated when comparing model performance.

3.2 ***Model Comparison***

Comparing the MSEs and visualizations of the models indicates that the best model was the linear regression model. The linear regression model had a lower testing MSE than the other models. Though the linear regression model was more conservative in its predictions than the gradient boosting regression model, the predictions of the linear regression model were still profound enough to consider the model useful.

Model	MSE
Linear Regression	14098
Random Forest Regression	16614
Gradient Boosting Regression	20792

Figure 13: MSE Table

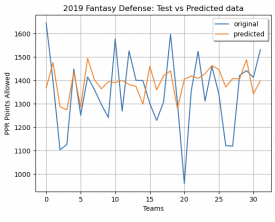
To compare the models further, the 2018 validation set was used to determine what predictions each model would make for the 2019 season. To compare these results,

the top ten defenses from the 2019 season according to FantasyPros.com are the following:

RANK	TEAM
1	New England Patriots
2	Pittsburgh Steelers
3	San Francisco 49ers
4	Baltimore Ravens
5	Kansas City Chiefs
6	Los Angeles Rams
7	Minnesota Vikings
8	New Orleans Saints
9	Buffalo Bills
10	Tampa Bay Buccaneers

Figure 14: FantasyPros.com Top Ten 2019 NFL Defenses

Figure 15 illustrates each model's results on the validation set. This includes the MSE, the predictions, the predicted top ten, and a comparison of the top ten predicted teams to the top ten results from FantasyPros.com. To note, the linear regression model maintained the lowest MSE. The majority of models predicted 4 out of 10 top defenses of the 2019 season.

Model Type (MSE)	Predictions	Top Ten Predictions	Matching Top Ten Teams to FantasyPros
Linear Regression (23421)		+----------------------+ | Tm| +----------------------+ | Buffalo Bills| | Minnesota Vikings| | Chicago Bears| | Baltimore Ravens| |Jacksonville Jaguars| | Tennessee Titans| |Los Angeles Chargers| | Dallas Cowboys| | Arizona Cardinals| | Pittsburgh Steelers|	Buffalo Bills, Minnesota Vikings, Baltimore Ravens, Pittsburgh Steelers

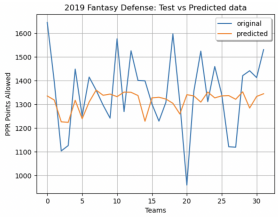
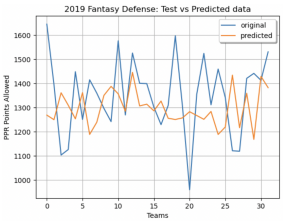
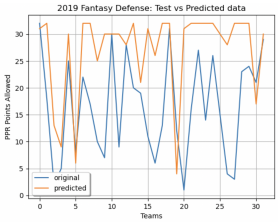
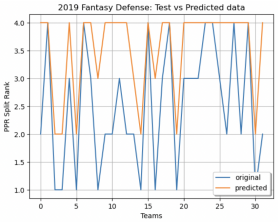
Random Forest Regression (23830)		+----------------------+ | Tm| +----------------------+ | Buffalo Bills| | Baltimore Ravens| |Jacksonville Jaguars| | Chicago Bears| | Minnesota Vikings| |Tampa Bay Buccaneers| | Miami Dolphins| | New York Giants| | Cincinnati Bengals| | Carolina Panthers|	Buffalo Bills, Baltimore Ravens, Minnesota Vikings, Tampa Bay Buccaneers
Gradient Boosting Regression (34073)		+----------------------+ | Tm| +----------------------+ |Tampa Bay Buccaneers| | Oakland Raiders| | Cincinnati Bengals| | San Francisco 49ers| | Philadelphia Eagles| | Cleveland Browns| | Atlanta Falcons| | Miami Dolphins| | New York Giants| | Carolina Panthers|	Tampa Bay Buccaneers, San Francisco 49ers
Random Forest Classifier - 32 (204)		+----------------------+----------+ | Tm|prediction| +----------------------+----------+ | Minnesota Vikings| 4.0| | Chicago Bears| 6.0| | Buffalo Bills| 9.0| | Baltimore Ravens| 13.0| | Tennessee Titans| 17.0| |Jacksonville Jaguars| 21.0| | Dallas Cowboys| 25.0| |Los Angeles Chargers| 26.0| | Houston Texans| 28.0| | Pittsburgh Steelers| 28.0|	Minnesota Vikings, Buffalo Bills, Baltimore Ravens, Pittsburgh Steelers
Random Forest Classifier - 4 (1.59)		+----------------------+----------+ | Tm|prediction| +----------------------+----------+ | Baltimore Ravens| 1.0| |Jacksonville Jaguars| 1.0| | Chicago Bears| 1.0| | Minnesota Vikings| 2.0| | Buffalo Bills| 2.0| | Indianapolis Colts| 3.0| |Los Angeles Chargers| 3.0| |New England Patriots| 3.0| | Philadelphia Eagles| 3.0| | Houston Texans| 3.0|	Baltimore Ravens, Minnesota Vikings, Buffalo Bills, New England Patriots

Figure 15: Model Predictions for 2019

3.3 Champion Model

After exploring the testing results and validation results of each model, we determined that the best model based on performance was the linear regression model. The model consistently outperformed the other models and maintained a lower MSE. With that being said, using the knowledge gained from the models collectively could be the best way to ensure a top defense is selected. For example if we cross reference the matching top ten teams of each model, we notice that three teams: the Buffalo Bills, Minnesota Vikings, and Baltimore Ravens are consistently predicted. Selecting one of these teams as your fantasy defense would result in having a top ten defense for the 2019 season.

4 Conclusions

4.1 Future Research

There are many options to explore when determining what future research could be conducted through the use of this project. The first is relatively obvious. Upon completion of the 2022 season, the defensive data could be collected and processed to make predictions for the 2023 season. The tricky part about this is that most fantasy data includes the total fantasy points scored for each defense instead of the total fantasy points allowed. To gain the data for total fantasy points allowed, the fantasy points allowed defensively for each skill position would need to be collected and combined. In a similar way, future research could use the results of this project to identify elite defenses and Super Bowl Champions. While exploring the data, we noticed that when we ordered the data by PPRDefRank and PPR_Points_Allowed all time defenses rose to the top of the ranks. Figure 16 shows the top ten defenses based on rank and PPR points allowed. To note, some of these defenses were elite! The 2002 Tampa Bay Buccaneers featured one of the baddest big men to ever play the game, Warren “QBK” Sapp, and they also won a Super Bowl. The 2004 and 2008 Pittsburgh Steelers included Hall of Famer Troy Polamalu and potential 2023 future Hall of Famer James Harrison. They won Super Bowls in 2005 and 2008. The 2006 and 2011 Baltimore Ravens included one of the best linebackers to ever play the game, Ray Lewis, as well as one of the best safeties to ever play the game, Ed Reed. Both players are Hall of Famers, and the Baltimore Ravens won the Super Bowl in 2012. The 2013 Seattle Seahawks featured one of the greatest secondaries ever assembled, the legion of boom. Furthermore, this team won a Super Bowl in 2014 and lost a Super Bowl in 2015.

Tm	Year	PPRFantasyPointsAllowed
Tampa Bay Buccaneers	2002	835.0
New York Jets	2009	845.06
Pittsburgh Steelers	2008	884.94
Pittsburgh Steelers	2004	919.8
Dallas Cowboys	2003	921.84
Baltimore Ravens	2006	931.94
New England Patriots	2019	959.84
Seattle Seahawks	2013	962.88
Chicago Bears	2005	964.78
Baltimore Ravens	2011	993.8

Figure 16: Data Ordered by PPRDefRank and PPRFantasyPointsAllowed

The results in Figure 16 show there is a clear correlation between the response variables we chose, elite defenses, and Super Bowl Champions. Thus, using this research to identify the next elite defense and potential Super Bowl Champion could be an excellent, and potentially lucrative, next step.

Another idea for future research could be to explore different metrics to compare the models. In this exploration, we used MSE as the metric to determine the best model. However, the MSE can sometimes be deceiving. To illustrate this idea, a logistic regression model was created to predict Next_PPRDefRank. The MSE of the model was 130. Based on MSE, the logistic regression model appears to be a better model than the random forest classifier. However, Figure 17 illustrates that the logistic regression model predicted the same value across the whole set. The improved MSE indicates a better model, but the visualization shows the model is clearly flawed. Thus, exploring metrics beyond MSE could produce a better understanding of the results.

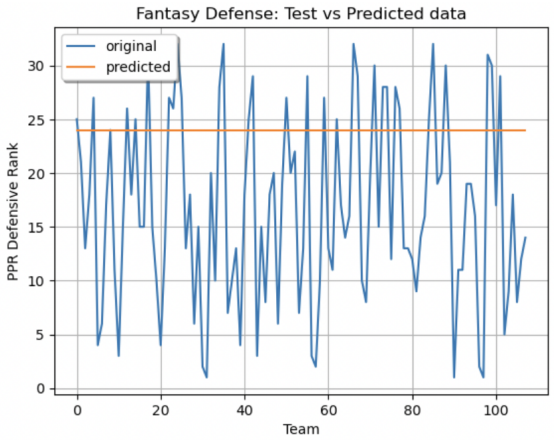


Figure 17: Logistic Regression Predictions

Incorporating offseason grades would also be an excellent next step as a future research idea. Each offseason, teams try to improve through the draft and free agency. By incorporating an offseason grade for each team, we could potentially improve model performance. Our group attempted to do this, but unfortunately ran into a number of issues in the data processing phase. First, we were only able to find offseason grades of seasons after 2016 as the data was not available before then. This was an issue because our data ranged from 2002 to 2017, meaning we would not be able to model with the variable. Another issue that caused concern during this process was the bias that could potentially be introduced from the offseason grades. When searching for offseason grades, we came across a large number of different opinions, grading systems, and rankings. In a sense, offseason grades themselves are a bit arbitrary. To gain the best result and to ensure the bias of a single evaluator would not be overemphasized, it would require a multitude of offseason grades for each team to be collected, standardized, and averaged to determine a more appropriate offseason grade for each team. Lastly, the third issue we ran into with the offseason grading system is that it did not account for player

aging and retirements, which leads to a decline in performance for the defense. This meant that teams still got worse even though they had what would be considered a great offseason according to the power rankings at the end of the previous season and the power rankings at the beginning of the following season. Figure 18 is a sample of the offseason grades that we manually constructed. Offseason grade values are represented on a 4.0 GPA scale. The data was based on information and averages from two prominent companies in the fantasy industry, pff.com and Bleacher Report.

NFL Team	Off Season Grade (2022)	Off Season Grade (2021)	Off Season Grade (2020)	Off Season Grade (2019)	Off Season Grade (2018)	Off Season Grade (2017)
Arizona Cardinals	2.333333333	3	3.666666667	3	2	3
Atlanta Falcons	3.333333333	2	2	1.333333333	2.666666667	3
Baltimore Ravens	4	3	3	2.333333333	2	2
Buffalo Bills	3.333333333	3	3.666666667	4	2.666666667	4
Carolina Panthers	3.333333333	2	2.666666667	3	1.333333333	3.666666667
Chicago Bears	3	3	2.333333333	1.666666667	4	2
Cincinnati Bengals	4	3.666666667	4	1	2.333333333	1
Cleveland Browns	3.333333333	4.333333333	3.666666667	1.666666667	4	2.333333333
Dallas Cowboys	2.333333333	2.333333333	3	2.666666667	1	3
Denver Broncos	4	2.666666667	3.333333333	1.333333333	3.333333333	4
Detroit Lions	3.666666667	3	2.666666667	2.333333333	3.666666667	2.666666667
Green Bay Packers	2.666666667	2.666666667	1.666666667	4	2	3.666666667
Houston Texans	3.666666667	1	1.333333333	2.333333333	2	4
Indianapolis Colts	3.333333333	1.333333333	3	4	3	2.333333333
Jacksonville Jaguars	2.333333333	2.666666667	3	3.333333333	0.6666666667	4
Kansas City Chiefs	3.333333333	3.666666667	2.333333333	2	2.333333333	3.666666667

Figure 18: Offseason Team Grades

Using strength of schedule as a predictor variable could also be an excellent direction for future research. We tried to incorporate strength of schedule as a predictor variable over the last two weeks as a final step to our project to try to improve our models. The theory was that, even if a defense was average, if the schedule was an easier one, a team would end up performing better than they should, making our model more robust. We found a dataset from FantasyDataPros that gave the strength of schedule and merged it with our dataset. The next step was to run the models on the new dataset. Figure 19 depicts the results.

Model	MSE
Linear Regression	14101
Random Forest Regression	15073
Gradient Boosting Regression	20447

Figure 19: MSE Table after Incorporating Strength of Schedule

We ended up getting better results for the random forest regression and the gradient boosting regression, but it did not end up giving us a better MSE for the linear regression, which was our champion model. Additionally, there were some errors that were brought up from merging the dataset that we were not able to solve when we tried to optimize our model more efficiently. Ultimately, we felt this job was only partially

finished, but the initial results were promising enough to include it in our future research ideas.

4.2 Summary

The aim of this exploration was to gain a deeper understanding of D/ST, and to determine if previous D/ST data can be used to predict future D/ST results. To accomplish our aim we used data from FantasyPros.com located on github. We processed the data by restricting the dataset to the years 2002 to 2017, and we engineered the response variables Next_PPR_Allowed and Next_PPRDefRank to predict each team's D/ST results of the next season. Using the response variables and predictors such as points allowed, TOs, and completions, we created linear regression, random forest regression, and gradient boosting regression models to predict Next_PPR_Allowed. We also created a random forest classifier model to predict Next_PPRDefRank. After tuning the models and comparing the results we determined that the best model based on performance was the linear regression model. The linear regression model had the lowest MSE at 14098, and predicted four top ten D/ST in 2019 using the 2018 validation set. Furthermore, a large number of our models predicted four top ten D/ST, and three teams in particular-- the Buffalo Bills, Minnesota Vikings, and Baltimore Ravens-- were common outcomes for most of the models. Each of these teams finished as top ten fantasy defenses in 2019. After comparing our models, we then decided to tune the models further by feature engineering the new variables offseason grade and strength of schedule. Although more research is needed to better incorporate these variables, we were able to improve the MSEs of both our random forest regression model and our gradient boosting regression model by including the strength of schedule variable as a predictor. Through this exploration we were able to gain a deeper understanding of D/ST, we strengthened our Spark skills through data collection, processing, and modeling, and we had a lot of fun exploring the great NFL defenses of the last two decades!

References

[1] ReportLinker (2022). *Fantasy Sports Market - Growth, Trends, COVID-19 Impact, and Forecasts (2022 - 2027),* Yahoo Entertainment, *www.yahoo.com/entertainment/fantasy-sports-market-growth-trends-154700504.html*

[2] Fox Sports (2022). *Fantasy NFL is a $70 billion industry — yes that's a 7 followed by ten zeros,* Fox Sports, www.foxsports.com.au/football/world-cup/fifa-world-cup-2022-portugal-vs-switzerland-live-updates-round-of-16-start-time-stats-score-blog-news-highlights/news-story/c9a21143aeb502d4b02e67d4245dcc3a?recommendedCount=0

Moton, M. (n.d.). *Bleacher report*. Bleacher Report. Retrieved December 11, 2022, from https://bleacherreport.com/

NFL, Fantasy Football, and NFL draft. PFF. (n.d.). Retrieved December 11, 2022, from https://www.pff.com/

Notifications. FantasyPros. (n.d.). Retrieved December 11, 2022, from https://www.fantasypros.com/nfl/reports/leaders/dst.php?year=2019

Wikimedia Foundation. (2022, October 28). *List of Super Bowl Champions*. Wikipedia. Retrieved December 11, 2022, from https://en.wikipedia.org/wiki/List_of_Super_Bowl_champions

Fantasydatapros. (n.d.). *Fantasydatapros/data: Fantasy football data in the form of CSV files available for use in pandas, R, Excel etc.*. GitHub. Retrieved December 12, 2022, from https://github.com/fantasydatapros/data