Predicting NFL Contracts Based on Player Statistics

Jack Arbuckle, Ian Pezzella, Ben Scartz, David Sobek, and Reuben Dayal

Introduction

This project looks to predict the salaries and contracts of current NFL players by looking at their previous year's statistics. This project uses a linear regression model, lasso regression model, random forest model, and XGBoost model to determine which model can predict former NFL contracts with the highest accuracy. The problem for this project is relevant to all football players and those who run NFL organizations. Being able to predict contracts is important for teams that invest a lot of money into the players themselves. It is also important for current NFL players to attempt to assess their worth. A player becoming a free agent can look at what contracts players with similar metrics around the league have received and deduce his own value.

The issue of player value and an organization determining a player's value is not new, and this analysis provides real-world analysis for this process. Many players are often considered under or overvalued for a variety of reasons, which can lead to pressure from the media and fanbases, salary cap issues, or potential front-office layoffs for making poor contract decisions. This project is looking to understand which on-field attributes weigh the most in terms of contract evaluation. This question is pondered by NFL executives and fans alike when looking at player contracts, and the goal is to accurately assess player value by seeing what on-field metrics carry the most weight within offensive NFL contracts.

Related Work

Many websites and data companies such as Spotrac [2] have come up with contract predictions based on market conditions, player statistics, and other factors. While many of these estimates are very detailed with outputs on contract length and amount, they fail to provide exact details on how they obtained these values. They fail to display the process and accuracy of their predictions.

Some previous studies have been conducted to try and solve a similar problem with salary prediction in the NFL. One study conducted by Martin Roach [3] used a player's marginal product of labor (MPL) to try and predict a player's next contract value. The MPL value is a rough estimate of how much a player is contributing to his team's success. This study took into account how much of the cap the position of the player in question takes up. The model took wins/losses and point differentials heavily into account. One of the biggest factors involved in this model was a variable called "salary out rate". This value quantifies the amount of player talent that is not used on the field throughout the year. The value that is generated represents the proportion of the salary cap that the position in question takes up. It was then broken down into different sides of the ball and further by different positions. It seeks to explain the

“value of resources lost to unexpected absences”. While this study does a good job of exploring the relationship between a player’s overall contributions to their team and their annual contract value, it is not clear what specific game statistics can throw off the model.

The project that was done for this report, sought to dive deeper into the specific statistical categories that affect the contract value of a player in various positions. Offensive skill positions such as quarterbacks, running backs, and wide receivers were the best positions to look at for the purposes of this study, as they have the most quantifiable statistics that can assess the level of their performance. While the previous study likely does a better job of predicting contract value at the macro level, our model likely does a better job of utilizing micro-level statistics.

Data Description

To gain data for this project, the package “nflfastr” was used to scrape play-by-play data from NFL games. [1] While Spotrac’s [2] website was used to extract NFL contract information. One of the challenges faced in compiling the data was that the data had to be merged so that the contract information and data were in one data frame. The nflfastr data included stats such as passing touchdowns, receiving yards after the catch, games played, fantasy points, etc.[1] Spotrac’s contract dataset includes stats such as signing bonus, guaranteed money, age, and length of contract.[2] Depending on what position is being analyzed, different statistics will need to be used. Summary statistics were taken to get an initial assessment of the data. The summary statistic that immediately stood out was the distribution of Average Annual Value (AAV). The median AAV for NFL players, according to this dataset, was $2.6 million. The mean AAV, however, was significantly higher, coming out to be about $6.6 million. This coupled with the fact that the maximum value here was $55 million indicates that this data is heavily skewed to the right. We also needed to scale the data because the units for the different variables all have different magnitudes. For example, rushing, passing, and receiving yards have a much wider range and higher max than stats such as fumbles and fumbles lost. The maximum number of yards is 1,653, while the maximum number of fumbles is 10. The data must be scaled accordingly, or else the stats with much larger raw values will be shown as better predictors.

Methods

Multiple Linear Regression

Multiple Linear Regression models the relationship between a dependent variable and two or more independent variables by fitting a linear equation to the data. For this project, multiple linear regression looks at all the imputed explanatory variables and attempts to precut the AAV of the player. The weakness

of linear regression is that it assumes a linear relationship where one may not be available. This could screw the data and fail to accurately predict the salaries of players that do not follow the linear trend.

Lasso

Lasso regression analyzes the importance of each independent variable by weighting each of them. The larger the correlation coefficient for each variable the higher the impact that variable has on predicting the response variable. The weakness of lasso regression is that like linear regression it assumes linearity, it struggles with many factors, and it is very sensitive to outliers in the data.

Random Forest

Random forest uses decision trees and then bagging to train multiple decision trees in the data set. One of the strengths of Random Foresting is that the decision trees are uncorrelated because at each split it only considers random features. Unlike some simpler models, Random Foreseting also does overfit to the training data. In R, the project focused on the node size for growing trees, the number of trees, and the number of predictors to use at each node, m.

XGBoost

In boosting, the trees are built sequentially in the model. Each tree was built based on the information from the previous tree. This means that each tree is fit based on a modified version of the original data. We combined the results of a large number of decision trees to produce our final predictions. The model learns slowly and may require many trees. There are a lot of parameters involved with XGBoost including eta, which controls the rate at which the model learns, and the number of interactions for each tree, which controls possible interactions in the tree. Some strengths of XGBoost include it has strong predictive power, its resilience to overfitting, and its sensitivity to outliers. Some negatives about XGBoost were that it was slow to train and hard to run in parallel since the trees were built in sequence.

We found that the XGBoost model had the best predictive value in this project.

Results

Regression Models

For our multiple linear regression, we split our data into testing and training sets for each position group and then applied an exhaustive search to determine the ideal predictors that would best fit our model. For our regression model of NFL quarterbacks, we were able to produce a model with 8 predictor variables and an adjusted R-squared of 0.8855496. By using this model we were able to predict QB's

average annual contract value with an error of roughly 8.5 million dollars. For our Wide Receiver model we obtained an R-squared value of 0.8342111 using 6 predictor values This model was a better predictor of contract value, being able to predict average annual value for receivers within 2.4 million dollars. Finally, our Running Back model used 6 predictor variables and resulted in an R-squared value of 0.8462455. This Linear Regression model was our best fit, as we were able to predict contract value within 1.9 million dollars.

QB

ME	RMSE	MAE	MPE	MAPE
-7147337	11863025	8611361	-130.2782	183.1083

RB

ME	RMSE	MAE	MPE	MAPE
433579.7	2270759	1930529	-32.73306	80.21121

WR

ME	RMSE	MAE	MPE	MAPE
-1229068	3008672	2372945	-83.14213	132.5215

Lasso

We created a Lasso model for each of the position groups in order to see which of the variables in our model was the most impactful to our overall results. For Quarterbacks, we found that the most impactful variables were passing touchdowns, fantasy points, passing first downs, and passing EPA. Our wide receiver model showed that the most important variables were receiving yards, receiving air yards, games played, and targets. Finally, the most impactful variables in our running back model were receptions, targets, receiving yards, and rushing touchdowns.

Random Forest

For our Random Forest models, we once again began our analysis by splitting our data into training and testing sets. After we separated our data, we ran a Random Forest Model on all three position groups which predicted the AAV using all variables. These models all used a total of 200 trees, had a node size of 1, and set our mtry value to 2. Our Random Forest model gave us much better results than our Linear Regression models. We were able to predict QB average annual contract value within 5.6 million dollars, Wide Receivers within 1.9 million dollars, and Running Backs within 1.8 million dollars.

QB

ME	RMSE	MAE	MPE	MAPE
-5455405	10480986	5688275	-104.9617	112.8679

RB

ME	RMSE	MAE	MPE	MAPE
923656.5	2580411	1715870	-12.92797	46.21488

WR

ME	RMSE	MAE	MPE	MAPE
-1201955	2658725	1899210	-90.59768	94.74645

XGBoost

The XG Boost Models were the models that gave us our best results. We set our models for each position group to have the same seed to ensure that each model had the same partitions of data. Our XG Boost models all had an eta of 0.05, they ran through 500 rounds of testing each, and we set our evaluation metric as a model error. Our Mean Average Error was very small for each of the three models. By using XG Boost, we were able to predict Quarterback contract values within $193,000, Running Back contract values within $506,000, and Wide Receiver contract values within $194,000.

QB

ME	RMSE	MAE	MPE	MAPE
125186.8	472101.4	193004.8	0.007954992	2.589531

RB

ME	RMSE	MAE	MPE	MAPE
441275.5	914421	506729.5	4.48977	8.4199

WR

ME	RMSE	MAE	MPE	MAPE
118110.7	461734.1	194171.4	0.6181631	4.073621

Discussion

The real world impact of the above models can help players and organizations decide the player value of free agents and players already on their roster. This project looks to determine what on field

factors are most important for quarterbacks, wide receivers, and running backs. From the above lasso regression, the variables passing touchdowns, fantasy points, passing first downs, and passing EPA are the best at predicting the AAV for quarterbacks. For wide receivers, the variables receiving yards, receiving air yards, games played, and targets are the best at predicting the AAV. Lastly, the variables receptions, targets, receiving yards, and rushing touchdowns looked to be the best at predicting the AAV for running backs. These variables were then put into the models to use as the explanatory variables. Although the random forest model produced predictions that were more accurate than the multiple linear regression, the XGBoost was the only model that predicted the AAV to a degree where it could be applied to the real world. For the linear regression model and the random forest model, an average error of under $5 million was considered "good," but in real-world scenarios, this error is too high to be applied to contract negotiations. Only the XGBoost model has a low enough error to be considered real-world applicable. All three models for the quarterbacks, wide receivers, and running backs had errors of less than a million dollars which was highly accurate for this project. For football and all professional sports, contracts are what make or break the team and the player. Championships are won based on good contract management while disappointing seasons are based on poor contracts. This model can help prevent both parties from being "cheated" by the stats that are presented. Although the model cannot be used for players with injuries that have no or limited statistics, this XGBoost model could be applied to all quarterbacks, wide receivers, and running backs with statistics from previous seasons.

Conclusion and Future Work

A lot of interesting trends and patterns were noticed from the output of the models. If this project was to be done again, a bigger dataset would be needed. Having more data to train and more data to evaluate would provide more accurate predictions for every single model. Also with more data, outliers could be seen/dealt with accordingly.

Also, including more factors than just player statistics could be used to be explanatory variables. It is common that many variables other than on-field statistics are used to evaluate what a player deserves.

Inflation would be a great variable that needs to be accounted for, as the market is growing exponentially, especially for the QB and WR positions. Organizational/team needs would also be good to add as these heavily impact the contracts that players receive.

Contributions

Jack: R Man, Discussion, (20%)

David: Results, Methods (20%)

Ian: Methods, Introduction (20%)

Ben: R Man, Methods (20%)

Reuben: Data Overview, Conclusion and Future Work (20%)

Bibliography

1. *An R package to quickly obtain clean and tidy NFL play by Play Data*. Dev status. (n.d.). https://www.nflfastr.com/
2. *NFL active player contracts*. Spotrac.com. (n.d.). https://www.spotrac.com/nfl/contracts/
3. Roach, M. A. (2018). Testing Labor Market Efficiency Across Position Groups in the NFL. Journal of Sports Economics, 19(8), 1093-1121. https://doi.org/10.1177/1527002517704021