1. **Introduction**

An emerging issue in this year's NFL offseason is that running backs (RBs) are not getting paid as much as other offensive skill positions such as wide receivers (WRs) and tight ends (TEs). Their roles in the games are mostly the same: carrying the ball, catching the ball, blocking opponents, and scoring. However, these days, the evolution of the game is more inclined towards passing rather than rushing. With the change in play style, the salaries of WRs and TEs have increased, while RBs have lost the market power they once had. The New York Times, Blatt said in 1997, Barry Sanders of the Detroit Lions was the second highest paid player in the league. Today, not a single RB is ranked in the top 100 (Blett, 2023). According to spotrac.com, the highest-paid running back in the league this year is Christian McCaffery, but he is ranked 111th among all positions. Also, Blatt wrote that since 2011, total pay for running backs and fullbacks has increased only about 11 percent. For every other offensive position, total pay has increased at least 90 percent (Blett, 2023). These days, quarterbacks (QB), WRs, and TEs also rush a lot in the game. According to Blett, in 2000, the top running back on each team was responsible, on average, for close to 60 percent of the team’s regular-season total rushing yards. However, in 2022, the teams’ top rushers accumulated only 47 percent of all rushing yards (Blett, 2023). Nevertheless, the salary of RBs is far less than that of any other offensive skill positions. In this project, I would like to create a new salary model to increase the value of RBs. Advanced stats like, Expected Points Average (EPA), Win Probability Added (WPA), Weighted Opportunity Rating (WOPR) are now used to evaluate player and these values affect on salary for players. However, Kim provides changes in performance based on contract design. As we see her journal, EPA and WPA is more weighted on receiving stats. The average WPA prior to and following a contract for RBs is 0.035, with an EPA of -3.35. For WRs, the average WPA is 0.52, and the average EPA is 17.85. TEs have an average WPA of 0.25 and an average EPA of 9.15 (Kim et al, 2018). Therefore, in this project, WPA and EPA are not the variables that make the value of RBs meaningful.

This project focuses on offensive skill positions’ salary for RB, TE, and WR. There are few sources that provide NFL play by play data. We could directly download file from pro-football-reference.com, through NFL.com application programming interface (API) or R package, called nflreadr. However, pro-football-reference.com does not have the advanced stats such as EPA, WPA, WOPR, etc. API does not provide information about which players are present on the playing field for each play, what formations are being used (aside from the “shotgun” formation), player locations, or pre-snap player movement (Urko et al., 2018). Compare to other two methods, nflreadr has both play by play data and salary data. And this library has standard stats and advanced stats.

Target consumers who are going to use this project’s result are mainly general managers and players. By using the new salary model, general managers get the validity to offer players a predicted salary. For players, they can negotiate when they make a new contract using new salary model.

1. **Methodology**

**2.1 Data Preprocessing**

In nflreadr play by play data, use load\_player\_stats() to get a players’ weekly stats. After, combine weekly stats into season stats. For this project, I used the 2015 to 2022 seasons’ data. For the salary data, the first data frame shows year\_signed, years, value, apy, guarteed etc. To use the same season data as play\_by\_play data, use unnest() functions to unnest a list-column that shows the detail of the year contracts. After obtaining the season salary for each player, combine the play-by-play data and salary together.

Players who sustain an injury at the beginning of the season have a low chance of receiving or carrying the ball. Therefore, for rushing attempts, assume that players attempt at least two rushes per game, and for receiving targets. Also, assume that players target at least one pass per game. Set a threshold of total rushing attempts over 30 and total receiving targets over 20.

Most of the evaluation models, such as WPA, EPA, Passing Air Conversion Ratio (PACR), and Dakota, which is the sum of Adjusted EPA and CPOE, are related to receiving stats. For the new model of salary, variables that have a high weight on receiving stats remove from the data frame. So, the variables that will be used in the project are shown in Table 1.

This new salary model tries to find RBs deserve more salary than given salary. Most of the stats that use in NFL are more related to receiving stats as shown in table 1. To assign a weight to the rusher, create a new variable that corresponds to the play ratio of each player on their respective teams. Most of NFL teams’ starting running backs have high ratio of offensive snaps than starting receivers share their targets. By using that concept, make new variables relate to paly ratio.

Given salary data frame shows total contract value and guaranteed value. Some players’ contract includes low guaranteed value with workout bonus, per game roster bonus and few other bonuses. But players who have great performance in game have a guaranteed value that covers a higher percentage of their contract value. Use Term Frequency (TF) – Inverse Document Frequency (IDF) to determine the how relevant those guaranteed value is to a given contract value (Pradeep.E, 2021). And for this project, it requires dividing each season’s salary by the number of contracts years. TF-IDF follows equation 1 below: value is the total contract value, guaranteed is the guaranteed value on his total contract value and contract year is the length of the contract.

|  |  |
| --- | --- |
| Term Frequency – Inverse Document Frequency | |
|  | (1) |

Overall, in the final data frame, there are 678 RBs, 1033 WRs, and 443 TEs. And the outcome variable is weighted average value per year.

Among the variables, the variables called racr and wopr represent advanced stats. RACR, which stands for Receiver Air Conversion Ratio, can be calculated using equation 2(a) below. WOPR, which stands for Weighted Opportunity Rating, can be calculated using equation 2(b) below.

|  |  |
| --- | --- |
| Advanced Stats | |
|  | (2) a |
|  | (2) b |

**Table 1. Variables for the new salary model**

|  |  |  |
| --- | --- | --- |
| **Variables** | | |
| first down pass | first down penalty rec | third down converted rec |
| third down failed rec | fourth down converted rec | fourth down failed rec |
| receiving yards | yards after catch | air yards |
| first down rush | first down penalty rush | third down converted rush |
| third down failed rush | fourth down converted rush | fourth down failed rush |
| rushing yards | carries | rushing tds |
| rushing fumbles | rushing fumbles lost | receptions |
| targets | receiving tds | receiving fumbles |
| receiving fumbles lost | receiving first downs | racr (receiver air conversion ratio) |
| target share | air yards share | Wopr (weighted opportunity rating) |
| Weight value | rush rate per team rush | pass rate per team pass |
| rush rate per team all play | pass rate per team all play | play rate per team all play |

* 1. **Methods**

In this project, all works are done by R programming. Packages that are mostly used are ‘tidyverse’, ‘tidymodels’, ‘plotly’ and ‘shinyapp’.

The 'tidyverse' is helpful when filtering, selecting, or grouping variables in a data frame. While data preprocessing, ‘tidyverse’ is easy to create new columns and order the data by a column.

The 'tidymodels' package is used for splitting/resampling data, preprocessing, and modeling. Due to the data having 2154 observations, split the data in to train and test split. 70% of data is used as train split. As all variables used in this project are numerical, therefore, in the preprocessing step, step\_Yeojohnson and step\_normalize are used for all predictors. For resampling step, use repeated k-fold cross-validation with 10 folds and repeats 5 times. For modeling part, as outcome variable is also numerical value, I used linear regression as base model. The base model is used because it shows the results intuitively and in a simple way. After the base model, Random Forest and XGBoost are employed to observe how the results differ from the base model. Additionally, the use of machine learning methods makes it easier to predict feature importance. Since a regression model is used in this project, the 'RMSE' is employed to evaluate the model. All the ‘RMSE’ results are discussed in chapter 3.

The ‘plotly’ and ‘shinyapp’ package are used for visualization. Components in the dashboard include salary comparison by positions and a comparison of each player's given salary with the predicted salary by the model. I will discuss the results in chapter 3 where the first plot demonstrates that salary comparison by position reveals how RBs receive lower salaries than any other positions. Even after the modeling, receiving stats have more impact on salary; therefore, the salary gap between the positions does not change dramatically. The second plot illustrates each player's given salary and predicted salary by the model. This plot shows whether players are overrated or underrated based on their stats.

1. **Results**

**Table2. RMSE values following each modeling method**

|  |  |  |
| --- | --- | --- |
| **Model** | **Train Split (Unit: Million)** | **Test Split (Unit: Million)** |
| Base Linear Regression | 3.64 | 3.99 |
| Random Forest | 3.78 | 4.00 |
| XGBoost | 3.74 | 3.98 |

**3.1 Model evaluation**

For a regression model, 'RMSE' is the most common metric used to evaluate the model. The outcome of this model is the salary value measured in millions of dollars. All models use the same train/test split data and k-fold cross-validation. The only difference among the models is the choice of modeling method. As table 2 shows the train split, base linear regression has the lowest RMSE value with 3.64. For the test split, the three models have almost the same RMSE, but XGBoost has the lowest RMSE at 3.98.

* 1. **Underpaid / Overpaid**

**Table 3. Percentage by Position of Overpaid or Underpaid**

|  |  |  |
| --- | --- | --- |
| **Position** | **Underpaid (%)** | **Overpaid (%)** |
| RB | 61.95 | 38.05 |
| TE | 53.95 | 46.05 |
| WR | 64.86 | 35.14 |

This Table 3 shows the percentage by position that is overpaid or underpaid. To compare the original salary and predicted salary, calculate the average for all predicted salaries from the models. As this table shows that about 62% of RBs, 54% of TEs, and 65% of WRs get underpaid. This result indicates that, according to the predicted salary model, over 50% of players are underpaid. I would like to identify the players and positions that experience the highest levels of overpayment and underpayment.

Table 4 displays the top 5 overpaid and underpaid players and positions, including their weighted original salary and the average predicted salary from the models.

**Table 4. Overpaid / Underpaid Players and Positions**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Players** | **Position** | **Weighted Salary** | **Avg Predicted Salary** | **Overpaid / Underpaid** |
| Mike Wellace | WR | 30.35 | 10.87 | Overpaid |
| DeAndre Hopkins | WR | 28.43 | 10.11 | Overpaid |
| Deebo Samuel | WR | 24.41 | 6.87 | Overpaid |
| Tyreek Hill | WR | 30.23 | 14.01 | Overpaid |
| DeAndre Hopkins | WR | 28.43 | 12.45 | Overpaid |
| Alshon Jeffery | WR | 1.32 | 10.33 | Underpaid |
| Michael Thomas | WR | 1.51 | 11.21 | Underpaid |
| Jarvis Landry | WR | 0.96 | 10.50 | Underpaid |
| Jarvis Landry | WR | 0.96 | 10.76 | Underpaid |
| Darnell Mooney | WR | 0.91 | 10.48 | Underpaid |

All top 5 underpaid or overpaid players hold the position of wide receiver. This result suggests that wide receivers often sign contracts with low salaries in their rookie seasons. When they deliver impressive performances, their predicted salaries increase, aligning with other higher-paid receivers. On the contrary, the reason for their overpayment may stem from signing substantial contracts with the team after their rookie deals. However, over time, their performance may not consistently reach outstanding levels, causing their statistics to regress to the mean. Consequently, their predicted salaries end up being lower than the contract value.

**3.3 Feature Importance / Accumulated Local Effect**

Using the 'IML' package in R helps find feature importance and generate accumulated local effect (ALE) plots for both Random Forest and XGBoost. To find the feature’s importance, I created a function to combine models and new data for prediction. After combining the two parameters, I was able to obtain predictions. After running the function, I was able to get feature importance for both models. Figure 1 displays the results of the feature importance for both models.

|  |
| --- |
|  |
| *(a)* |
|  |
| *(b)* |

*Figure 1. Feature Importance for the models : (a) Random Forest model, (b) XGBoost model*

Figure 1(a) shows the feature importance for the Random Forest model. The top three important features in this model are WOPR, target share, and first down pass. Most of the rushing stats have less importance in the model. Figure 1(b) shows the feature importance for the XGBoost model. The top three important features in this model are WOPR, first down pass and target share. Also, most of the rushing stats have less importance in the model. As shown in the feature importance plots, new model for the salary impacted more on receiving stats, especially on WOPR, first down pass and target share. By feature importance plot, it says that rushing stats mostly doesn’t affect the salary model.

After plotting feature importance, draw a plot of ALE using the most important feature which is WOPR. Figure 2 shows the ALE plots for both models.

|  |
| --- |
|  |
| *(a)* |
|  |
| *(b)* |

*Figure 2. ALE plots for the models : (a) Random Forest model, (b) XGBoost model*

Figure 2 displays the ALE plots, illustrating the differences in predictions based on conditional distributions. The plots follow the most important feature, which is WOPR, in both the Random Forest and XGBoost models. For both plots, they demonstrate a similar relationship between weight value and WOPR. For both models, the ALE plot shows an exponential increase as WOPR increases. This indicates that WOPR has a significant impact on the predicted weight value, and its impact becomes more pronounced as WOPR increases.

* 1. **Shiny App**

|  |
| --- |
|  |
| *(a)* |
|  |
| *(b)* |

*Figure 3. (a) C*omparison of ea*ch player's given salary with the predicted salary by the model and (b) salary comparison by positions.*

As discussed in chapter 2, final visualization was used by shinyapp. Figure 3 shows the visualization results of this project. Figure 3(a) illustrates how the salary differs across each position. Even after the modeling, all three positions have a similar average salary. All units for salary are in million dollars. The average original salary for RBs was 2.37, for TEs it was 3.52, and for WRs, it was 4.40. After base linear regression model, all positions’ average salary has increased. The average salary from linear regression is 2.67 for RBs, 3.65 for TEs, and 4.60 for WRs. The average salary from Random Forest is 2.85 for RBs, 3.69 for TEs, and 4.67 for WRs. Lastly, the average salary from XGBoost is 2.66 for RBs, 3.80 for TEs, and 4.54 for WRs. After modeling, the average salary for all positions has increased. Among all the models, the largest increase observed was for RBs. This result covers all the players, so the results may differ for individual players. Figure 3 (b) shows the salary comparison for individual players. In the select option, we can choose seasons and players. Shown in Figure 3 (b) is Antonio Gates’ 2015 season salary. His original salary is above 6 million dollars, but through modeling, his salary decreases in all three models.

By using this final project, general managers in each team can negotiate with the players when they make a new contract. Following the variables that we used in this project, they can compare the players' stats with the value of salary. Also, players can negotiate with general managers to secure a higher salary than initially offered.

1. **Discussion/Conclusion**

The goal of this project is to advocate for higher salaries for running backs. To achieve this goal, I obtained play-by-play data and salary data for offensive skill position players from the 2015 to 2022 seasons. Considering that some players may get injured at the beginning of the season or during the offseason, and there may be few backup players, I established thresholds for rushing attempts and targets. As most of the new evaluation models, such as EPA and WPA, give high weights to receiving stats, I created a play ratio for each player. By using this variable, rush plays get higher weights than receiving plays. For the salary data, some players have fully guaranteed contracts, where the guaranteed value equals the contract value, while others receive bonuses in addition to the guaranteed value. To account for that difference, I created a new variable, 'weightvalue,' using the TF-IDF algorithm.

After completing data preprocessing, I split the data into training and testing sets, with a 70% training split. Then, I ran baseline linear regression, Random Forest, and XGBoost models. From those models, the RMSE values for the training split data were 3.64, 3.78, and 3.74 in sequence. And for the testing set were 3.99, 4.00, and 3.98. As those three models have almost the same RMSE, it provides a standardized measure that indicates their performance is similar. Based on the feature importance analysis, both Random Forest and XGBoost indicate that WOPR is the most important feature in these models.

After modeling, I used ShinyApp to check whether the average salary by positions differs from the original salary and how the average salary for each player has changed after modeling. Average salary by position stays mostly the same as original salary. As discussed in chapter 3, the average salary for all positions increased after modeling, especially, the average salary for RBs increased the most. The average salary for each player differs based on their stats.

Limitation of this project is first of all, receiving stats impact the most on salary decision. Therefore, players with high rushing stats and low receiving stats receive a lower salary than players with the opposite stats. Additionally, in the beginning, most WR and TE had higher salaries than RB, so the salary model places more emphasis on receiving stats than rushing stats. The second limitation of this project is that existing player evaluation models like EPA and WPA change significantly with a high ratio of emphasis on receiving. There is not much ongoing research on a rushing evaluation model in the NFL. Even though the NFL is more of a passing game than a rushing game, rushing still has a significant impact on the game.

For future research, I aim to develop a new evaluation model that places a high emphasis on rushing plays. Additionally, in the 4th quarter when there are less than 2 minutes remaining in a winning situation, rushing plays are useful for running down the clock with minimal risk compared to passing plays. Hence, for future research, I aim to explore the impact of the remaining time on the game clock and situational factors on the efficiency of plays. Utilizing this efficiency data, I plan to develop a new evaluation model for both rushing and passing plays. Using the new evaluation model, I intend to create a salary model similar to the one developed for this project. I will then compare how it differs from the actual salary paid and the predicted value obtained without incorporating the new evaluation model.

**References**

Blatt, B. (2023, September 7). The Plight of the N.F.L. Running Back. The Upshot, The New York Times. Retrieved from https://www.nytimes.com/2023/09/07/upshot/nfl-running-backs-pay.html

Kim, S., Sarina, A., & Sarin, S. (2018). Do players perform for pay? An empirical examination via NFL players’ compensation contracts. Journal of Banking & Finance, 88, 330–346.  
doi:10.1016/j.jbankfin.2018.01.004

Pradeep, E. (2021, March 21). Understanding TF-IDF in NLP: A Comprehensive Guide. Medium. https://medium.com/@er.iit.pradeep09/understanding-tf-idf-in-nlp-a-comprehensive-guide-26707db0cec5

Spotrac. (n.d.). https://www.spotrac.com/nfl/rankings/contract-value/

Urko, R., Ventura, S., & Horowitz, M. (2018). nflWAR: A Reproducible Method for Offensive Player Evaluation in Football (Extended Edition). Department of Statistics & Data Science, Carnegie Mellon University. Retrieved from https://arxiv.org/pdf/1802.00998.pdf