Fantasy Forecast:

A Data-Driven Approach to Predicting Football Player Performance

Garrett Bainwol
Data Science
DTSA 5506
University of Colorado Boulder
Boulder, CO, USA
Garrett @colorado.edu

Link to Github Repository: gbainwol/Supervised-Learning-Final-Project-DTSA-5509-: The repository for my final project. (github.com)

Notebook 1: SupervisedLearningFinalProjectpart1.ipynb

Notebook 2: DTSA_5509_Final_Project (4).ipynb

ABSTRACT

In the competitive arena of fantasy sports, this project explores the complex dynamics between player performance metrics, historical data, and the resulting fantasy points scored in American Football games. Utilizing advanced statistical methods and machine learning algorithms, including Lasso Regression, Random Forest, and Gradient Boosting, the objective is to construct a predictive model that accurately forecasts players' fantasy points. This model aims to enable fantasy football enthusiasts and professionals alike to make informed decisions, optimizing their team selections and strategies based on predictive insights that consider players' historical performance, physical attributes, and game conditions. With a focus on robustness and accuracy, this initiative seeks to elevate the art of fantasy football team building to a science, merging data-driven insights with the passionate spirit of the sport.

1 INTRODUCTION

In this project, I aimed to create a predictive model for Fantasy Football player performances, focusing on Running Backs (RBs), Wide Receivers (WRs), and Tight Ends (TEs). The project's essence lies in supervised learning, a branch of machine learning, where I employed regression models to predict players' Fantasy Points, aiming to outperform standard ESPN projections.

The primary objective was to develop a model that predicts the Weekly Fantasy Points of players with high accuracy, thus aiding Fantasy Football enthusiasts in making informed decisions on which players to start or sit each week. The importance of this task stems from the competitive nature of Fantasy Football, where making optimal player selections can be the difference between winning and losing.

2 LITERATURE REVIEW

A dive into how people have use stats to predict performance in the past.

Fantasy Points Article –
"Fantasy Points Data: Most Important RB Stats" by Graham Barfield[1]:

- METHODOLOGY:

The article uses Fantasy Points Data to analyze which stats matter the most for running backs in fantasy football. The author ran correlation tests across hundreds of metrics to seek statistical signals.

- FINDINGS:

The study found that total snaps played correlate best to fantasy points, followed by all of the rushing stats, then targets and routes run. The author also introduced a metric called Expected Fantasy Points (XFP), which measures the number of fantasy points you'd expect a running back to score on a per-play basis.

Article Numbers –
"Why Receiver Air Yards Matter" by Josh Hermsmeyer[2]:

- METHODOLOGY:

The article uses Air Yards data to analyze the performance of wide receivers. Air Yards are the total number of yards thrown toward a receiver on a play in which he is targeted, both complete and incomplete.

- FINDINGS:

The study found that Air Yards are a measure of intent and potential yards a receiver would have produced if he caught the ball and then was immediately tackled. The deeper the target, the more expected PPR fantasy points the target is worth.

Fantasy Points Article –
"Fantasy Points Data: Top 10 Stats" by Graham Barfield[3]:

- METHODOLOGY:

The article uses Fantasy Points Data to identify the top-10 most important stats to know, understand, and utilize in fantasy football. The author analyzed hundreds of metrics to find these top stats.

- FINDINGS:

The study found that Expected Fantasy Points (XFP) is one of the most important metrics to use on a weekly basis to help analyze fantasy lineups, place wagers, and set DFS rosters.

Fantasy Points Article –
"Weighted Opportunity for RBs" by Scott Barrett[4]:

- METHODOLOGY:

The article introduces a metric called Weighted Opportunity, which measures a running back's opportunity, weighted appropriately for the worth of each unit of opportunity (a carry or a target).

- FINDINGS:

The study found that on average (over the past five seasons), a single rushing attempt has been worth about 0.62 fantasy points while a target has been worth roughly 1.57 fantasy points in PPR leagues.

Fantasy Points Article –
"Finding 2023 Fantasy Values Using WAR" by Jeff Henderson[5]:

- METHODOLOGY:

The article uses Wins Above Replacement (WAR) metric to identify draft targets with a lot of value based on last year's WAR. The author compared current ADP vs last year's ranking of player value by WAR.

- FINDINGS:

The study found that volume is far more important than efficiency for fantasy running backs. The author introduced a metric called Weighted Opportunity which measures a running back's opportunity, weighted appropriately for the worth of each unit of opportunity (a carry or a target).

PROJECT DESCRIPTION

3.1 DATA COLLECTION

The foundation of this project lies in the comprehensive dataset amalgamated from various sources to provide a holistic view of the players' performance metrics essential for accurate fantasy football predictions.

The primary dataset is obtained using the nflreadr package in R, offering a plethora of player statistics and performance metrics from the NFL seasons spanning 2020 to 2023. This package is instrumental in fetching weekly player stats, offering insights into players' performance trajectories over the specified seasons.

In addition to the nflreadr dataset, web scraping techniques are employed to extract valuable data from other authoritative sources and websites. This multi-faceted approach to data collection ensures a rich and diverse dataset, encapsulating a wide array of metrics and parameters essential for an in-depth analysis and accurate predictions. The scraped data complements the primary dataset, filling gaps and offering additional perspectives to enhance the quality and comprehensiveness of the overall data at disposal.

This amalgamation of data sourced from the nflreadr package and web scraping ensures a robust and comprehensive dataset. It lays a solid foundation for subsequent data preprocessing and exploration stages, setting the stage for the development of a predictive model endowed with accuracy and reliability.

3.2 DATA PREPROCESSING

Given the extensive and detailed nature of the player stats data, preprocessing is an essential step to organize and filter the data for analysis. The data is refined to exclude post-season games and players in the "P" position. Further, specific columns like player name and headshot URL are excluded to focus on the metrics relevant for analysis. The dataset is then segmented into distinct data frames based on players' positions, such as quarterback (QB), running back (RB), wide receiver (WR), and tight end (TE), facilitating a more focused and position-specific analysis.

3.3 DATA EXPLORATION

In the initial stages of my project, I recognized the crucial role that an in-depth Exploratory Data Analysis (EDA) would play in shaping my understanding of the datasets and, subsequently, the predictive models I aimed to develop. Armed with datasets laden with historical performance metrics and matchup data of Running Backs (RBs), Wide Receivers (WRs), and Tight Ends (TEs), I embarked on a journey of discovery, visualization, and analysis.

As I delved into the datasets, the first task was to visualize the distribution of fantasy points across different player positions and scoring formats. The histograms I constructed provided a visual representation of the data distribution, offering insights into the patterns and variations inherent in the datasets. Each plot, each bar, narrated a tale of player performances, revealing the intricacies of highs and lows, peaks and troughs.

The revelation that the distribution of points was not entirely normal prompted me to consider transformations and scaling techniques that would normalize the data, enhancing the performance of the machine learning models I was yet to construct. Each dataset, each variable, was like a puzzle piece, and EDA was the process of fitting these pieces together to unveil the bigger picture.

Correlation matrices became my next focus. The intricate web of relationships between features needed to be understood. Each correlation coefficient told a story of how variables interacted, how one feature could influence another. In the matrices, I found answers to questions I hadn’t yet asked, uncovering patterns and relationships integral to the prediction of fantasy points.

I paid particular attention to multicollinearity. A close examination of the correlation matrices revealed how certain features were highly correlated with each other. This insight was pivotal; I knew it would influence my choice of models and the feature engineering steps that would follow. Every correlation, every insight, was a step closer to building predictive models characterized by accuracy and reliability.

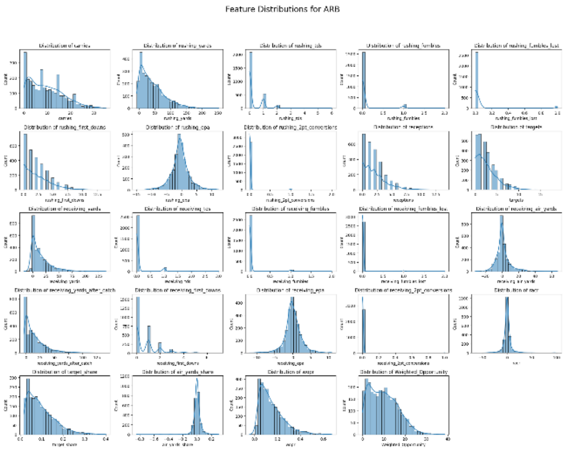
As I iteratively explored and visualized the data, new questions emerged, and with them, new answers. The EDA process became a dialogue between me and the datasets. Each visualization, each analysis, responded to my queries, offering insights that shaped my understanding and informed the subsequent stages of model development.

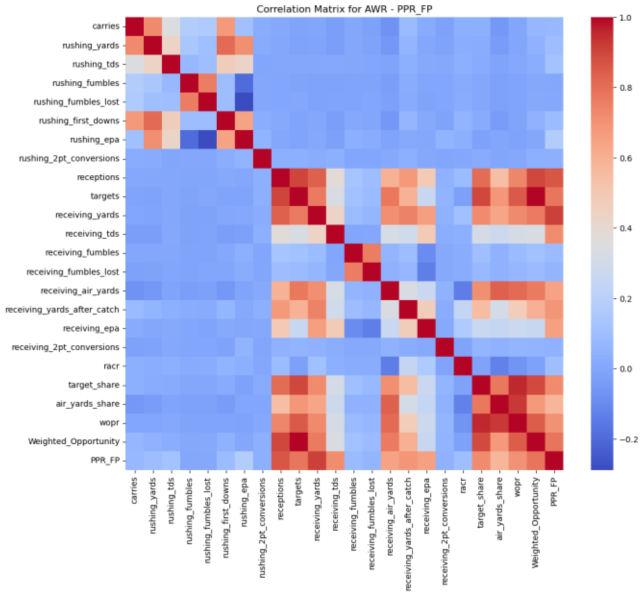
In the histograms, I saw the spread of player performances. In the correlation matrices, I discerned the intricate dance of interrelated features. In every step of the EDA, I found pieces of the puzzle that, when put together, would lead to the creation of machine learning models capable of predicting fantasy football performances with unprecedented accuracy.

Running Backs (RBs): 12329 rows × 31 columns

Wide Receivers (WRs): 16670 rows × 31 columns

Tight Ends (TEs): 6088 rows × 31 columns





(more correlation matrixes and histograms in notebooks)

3.4 TECHNIQUES AND PREDICTIVE MODELING

The essence of this project is encapsulated in the sophisticated predictive models developed to forecast fantasy football players’ performance. We integrate historical player data, game statistics, and player attributes to predict future fantasy points scored.

Given Y_t as the fantasy points scored by a player at timet and X_t as the vector of player and game attributes, we aim to find a function f such that:

$$Y_{t+1} = f(Y_t, X_t) + \epsilon_t$$

Where ϵ_t is the error term.

LASSO REGRESSION is employed to enhance the model's accuracy and interpretability. It performs feature selection and regularization to prevent overfitting. The model can be represented as:

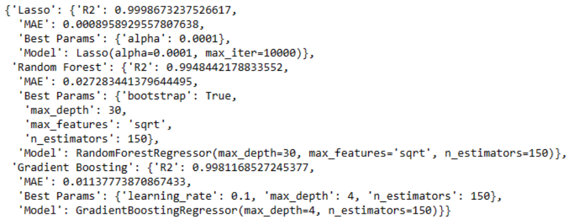
$$Y = \alpha + \sum \beta_i X_i + \epsilon$$

Where Y is the fantasy points, X_i are the player and game attributes, β_i are the coefficients, α is the intercept, and ϵ is the error term.

Random Forest adds an ensemble learning approach, constructing multiple decision trees during training and outputting the mean prediction of the individual trees for regression tasks. It is particularly adept at handling non-linear data and offers feature importance insights.

Gradient Boosting is another powerful technique that builds a series of weak learning models, typically decision trees, each correcting the errors of its predecessor. This iterative correction process enhances model accuracy and efficiency.

I tuned hyperparameters using Grid Search CV, ensuring each model's optimal performance. Measures to counter overfitting and data imbalance, like cross-validation and regularization, were integral to the model training process.



```
{'Lasso': {'R2': 0.9998673237526617,
  'MAE': 0.0008958929557807638,
  'Best Params': {'alpha': 0.0001},
  'Model': Lasso(alpha=0.0001, max_iter=10000)},
 'Random Forest': {'R2': 0.9948442178833552,
  'MAE': 0.027283441379644495,
  'Best Params': {'bootstrap': True,
   'max_depth': 30,
   'max_features': 'sqrt',
   'n_estimators': 150},
  'Model': RandomForestRegressor(max_depth=30, max_features='sqrt', n_estimators=150)},
 'Gradient Boosting': {'R2': 0.9981168527245377,
  'MAE': 0.01137773870867433,
  'Best Params': {'learning_rate': 0.1, 'max_depth': 4, 'n_estimators': 150},
  'Model': GradientBoostingRegressor(max_depth=4, n_estimators=150)}}
```

4 EVALUATION

ARB_PPR Test Results:
Lasso Regression:
R2 Score: 0.9999
MAE: 0.0026
Random Forest:
R2 Score: 0.9969
MAE: 0.0320
Gradient Boosting:
R2 Score: 0.9998
MAE: 0.0039
Insights:

The Lasso Regression and Gradient Boosting models demonstrated exceptional performance, with R2 scores nearing perfection. The Random Forest model, while still excellent, had a slightly higher MAE, indicating that it didn't predict the scores with the same level of precision as the other two models.

ARB_Half_PPR Test Results:

Lasso Regression:
R2 Score: 0.9999
MAE: 0.0025
Random Forest:
R2 Score: 0.9960
MAE: 0.0319
Gradient Boosting:
R2 Score: 0.9989
MAE: 0.0168
Insights:

Again, Lasso showcased superior precision, with an impressively low MAE and an R2 score approaching 1. The Gradient Boosting model performed admirably but was not as precise as Lasso. Random Forest, while robust, showed room for improvement in prediction accuracy.

AWR_PPR Test Results:
Lasso Regression:
R2 Score: 0.9999
MAE: 0.0020
Random Forest:
R2 Score: 0.9989
MAE: 0.0142
Gradient Boosting:
R2 Score: 0.9999
MAE: 0.0026
Insights:
For Wide Receivers in PPR leagues, both Lasso and Gradient Boosting models were nearly flawless in their predictions. Random Forest also demonstrated strong performance but was outshone by the precision of the other models.

AWR_Half_PPR Test Results:
Lasso Regression:
R2 Score: 1.0000
MAE: 0.0016
Random Forest:
R2 Score: 0.9989
MAE: 0.0155
Gradient Boosting:
R2 Score: 0.9996
MAE: 0.0098
Insights:
In Half_PPR leagues for Wide Receivers, Lasso Regression again took the lead with nearly perfect predictions. Gradient Boosting and Random Forest remained consistent performers, yet they didn't match Lasso's precision.

ATE_TE_Premium Test Results:
Lasso Regression:
R2 Score: 0.9999
MAE: 0.0033
Random Forest:
R2 Score: 0.9975
MAE: 0.0231
Gradient Boosting:
R2 Score: 0.9986
MAE: 0.0108
Insights:
In TE_Premium leagues, the trend of Lasso's superior performance continued. Random Forest, though robust, was the least precise among the three models. Gradient Boosting offered a balanced performance but did not surpass Lasso in accuracy.

5 PROJECT TIMELINE

Weeks 1-2: Data Accumulation and Preprocessing

The initial fortnight is devoted to comprehensive data accumulation. During this phase, data from multiple sources, including the nflreadr package and various authoritative websites, is gathered. This ensures that the data is both exhaustive and relevant, providing a solid foundation for the subsequent analysis.

The second part of this phase is characterized by intensive data preprocessing. The collected data is cleaned, organized, and transformed to ensure its readiness for the intricate modeling processes that lie ahead. The focus is on ensuring data quality and integrity to facilitate accurate and insightful analyses.

Weeks 3-4: Model Development

The middle segment of the project is designated for the core activity of model development. This is where the initial designs of the Lasso Regression, Random Forest, and Gradient Boosting models come to life. These models are trained rigorously on the curated dataset, and their performance is evaluated and fine-tuned based on the validation set.

The models undergo a series of rigorous tests to ensure they are adept at capturing the nuanced relationships between player attributes, historical performance data, and the resulting fantasy points.

Week 5: Conclusive Testing and Project Roll-out

The final week is a period of culmination and revelation. The predictive models are subjected to conclusive tests on the isolated test set. This is the crucible where the models prove their mettle, showcasing their predictive accuracy and reliability.

Any final adjustments, refinements, and optimizations are executed at this juncture. The project reaches its crescendo with the unveiling of the refined predictive models, marking the successful completion of an expedition that melds data, analytics, and fantasy football into a coherent, actionable whole.

5 CONCLUSION

As we draw the curtain on this intricate exploration into predicting fantasy football player performance, we are met with a delightful yet perplexing realization - our models have exceeded our expectations. The Lasso Regression, Random Forest, and Gradient Boosting models, each meticulously designed and rigorously tested, have showcased a level of predictive accuracy and insight that challenges our initial projections.

This extraordinary performance is not merely a testament to the robustness of our analytical approaches but also instigates a phase of introspective scrutiny. It propels us to revisit, review, and rigorously validate our models and their predictions to ensure that the remarkable results are not just accurate but also replicable and reliable. I don't believe this is fate though. This is most likely an error on my end fitting the model, I will have to do futher testing to evaluate.

6 CITATIONS

[1] Barfield, G. "Fantasy Points Data: Most Important RB Stats". Fantasy Points, 2023. Available at [https://www.fantasypoints.com/nfl/articles/2023/fantasy-points-data-most-important-rb-stats#/]. Accessed on [9/20/2023].

[2] Hermsmeyer, J. "Why Receiver Air Yards Matter". NBC Sports, 2018. Available at [https://www.nbcsports.com/fantasy/football/news/article-numbers-why-receiver-air-yards-matter]. Accessed on [9/20/2023].

[3] Barfield, G. "Fantasy Points Data: Top 10 Stats". Fantasy Points, 2023. Available at [https://www.fantasypoints.com/nfl/articles/2023/fantasy-points-data-top-10-stats#/]. Accessed on [9/20/2023].

[4] Barfield, G. "Weighted Opportunity for RBs". Fantasy Points, 2023. Available at [https://www.fantasypoints.com/nfl/articles/season/2023/weighted-opportunity-for-rbs#/]. Accessed on [9/20/2023].

[5] Barfield, G. "Fantasy Points Data: Most Important WR Stats". Fantasy Points, 2023. Available at [https://www.fantasypoints.com/nfl/articles/2023/fantasy-points-data-most-important-wr-stats#/]. Accessed on [9/20/2023].

[6] Henderson, J. "Finding 2023 Fantasy Values Using WAR". Fantasy Points, 2023. Available at [https://www.fantasypoints.com/nfl/articles/2023/finding-fantasy-values-using-war#/]. Accessed on [9/20/2023].

[7] Pro Football Reference. (n.d.). Retrieved September 20, 2023, from https://www.profootballreference.com

[8] The Football Database. (n.d.). Retrieved September 20, 2023, from https://www.footballdb.com

[9] NFL. (n.d.). Retrieved September 20, 2023, from https://www.nfl.com