Relationship Between Rookie Wide Receiver Stats and Future NFL Production

Glenn P. Dalbey

Western Governors University

Table of Contents

[A. Project Highlights 4](#_Toc177423276)

[B. Project Execution 5](#_Toc177423277)

[C. Data Collection Process 7](#_Toc177423278)

[C.1 Advantages and Limitations of Data Set 7](#_Toc177423279)

[D. Data Extraction and Preparation 8](#_Toc177423280)

[E. Data Analysis Process 9](#_Toc177423281)

[E.1 Data Analysis Methods 9](#_Toc177423282)

[E.2 Advantages and Limitations of Tools and Techniques 10](#_Toc177423283)

[E.3 Application of Analytical Methods 11](#_Toc177423284)

[F Data Analysis Results 11](#_Toc177423285)

[F.1 Statistical Significance 11](#_Toc177423286)

[F.2 Practical Significance 12](#_Toc177423287)

[F.3 Overall Success 13](#_Toc177423288)

[G. Conclusion 13](#_Toc177423289)

[G.1 Summary of Conclusions 13](#_Toc177423290)

[G.2 Effective Storytelling 13](#_Toc177423291)

[G.3 Recommended Courses of Action 14](#_Toc177423292)

[H Panopto Presentation 15](#_Toc177423293)

[References 16](#_Toc177423294)

# Project Highlights

**Research Question or Organizational Need:**

Can we predict future NFL wide receiver production by breaking down and evaluating statistics from their rookie season? Moreover, can this be used to create an advantage in fantasy football?

**Scope of the Project:**

* Included in Project Scope:
  + Data Collection and Cleaning
  + Statistical Analysis
  + Model development
  + Model validation
  + Visualization
* Not included in Project Scope:
  + Primary Collection of Data
  + Future data updates

**Tools and methodologies:**

* Methodology: CRISP-DM
* Environments:
  + Python
    - Pandas, NumPy, Matplotlib, Scikit-learn, SciPy
  + Jupyter Notebooks
  + Excel
* Analysis: ML Gradient booster

The use of these tools allowed for a complete analysis that provided an in-depth look at the metrics involved to help project future values of NFL wide receivers' fantasy value. This can benefit fantasy football players by helping them determine draft value and player forecasts.

# Project Execution

The execution of this project followed the initial plan outlined in the Task 2 document.

**B1. Goals, Objectives, and Deliverables**

The project achieved success by adhering to the project plan and executing the objectives as laid out. The following objectives and deliverables were successfully completed:

* + Objective 1: Obtain clean data for rookies and regular season wide receiver stats from 2006 – 2023.
    - Deliverable 1.1: Clean and process rookie data and years into a data frame.
    - Deliverable 1.2: Clean and process all wide receiver data by years with over one thousand-yard seasons into a data frame.
    - Deliverable 1.3: Combine cleaned data into a data frame containing all stats.
  + Objective 2: Obtain and clean data for the 2023 rookie season.
    - Deliverable 2.1: Cleaned and processed 2023 rookie stats in the data frame for analysis to be performed later.
  + Objective 3: Analyze data for correlations to determine the most relevant statistics to be used for machine learning.
    - Deliverable 3.1: Create a new data frame with the most relevant data for prediction models.
  + Objective 4: Create an ML Gradient Booster model to predict wide receiver future production.
    - Deliverable 4.1: Model creation with an accuracy score higher than 80%
    - Deliverable 4.2: The model is saved and will be used for real-world data to forecast production for 2023 and future NFL wide receivers.
  + Objective 5: Create a visualization that can be easily understood and able to communicate findings.
    - Deliverable 5.1: Visualizations that show past and future predictions.

**B2. Scope of the Project:**

The project scope followed what was planned and included the following elements:

* Included in Project Scope:
  + Data Collection and Cleaning
  + Statistical Analysis
  + Model development
  + Model validation
  + Visualization

**B3. Standard methodology**

The project used the CRISP-DM to follow and help guide the process. It ensured that the project stayed on track throughout and was executed in a structured manner.

**B4. Timeline and Milestones**

The project timelines precisely followed the planned dates and milestones.

|  |  |  |  |
| --- | --- | --- | --- |
| **Milestone** | **Duration** | **Start Date** | **End Date** |
| Data Collection | 5 Days | 8/12/24 | 8/16/24 |
| Data Cleaning | 2 Days | 8/19/24 | 8/20/24 |
| Data Analysis | 1 Day | 8/21/24 | 8/21/24 |
| Data Preparation | 1 Day | 8/22/24 | 8/22/24 |
| Data Integration | 1 Day | 8/23/24 | 8/23/24 |
| Statistical Analysis | 1 Day | 8/24/24 | 8/24/24 |
| Model Development | 3 Days | 9/3/24 | 9/6/24 |
| Model Evaluation | 1 Day | 9/10/24 | 9/10/24 |
| Real-World Report  Evaluation | 1 Day | 9/11/24 | 9/11/24 |
| Visualization and Creation | 1 Day | 9/12/24 | 9/12/24 |

The planned timeline was consistent because of planned downtime and accounting for problems along the way.

# Data Collection Process

## C.1 Advantages and Limitations of Data Set

**Advantages:**

* + The reliability of the datasets from professional pay websites allowed for the use of in-depth statistics and metrics.
  + CSV file formats created an easy transition to work with.
  + Considerable number of sources to collect data from, allowed to get all data needed even if not available from one site.

**Disadvantages:**

* + We were unable to get all the data from one source, so we had to use multiple sources to collect data, which led to additional transformations and cleaning.
  + Not all data was in CSV format, so it had to be converted.
  + Some metrics that had been planned to be used needed to be dated further and ruled out.

In general, the data used allowed for in-depth analysis and provided the metrics needed to create a model that successfully predicted future wide receiver production with reasonable accuracy.

# D. Data Extraction and Preparation

The data extraction process used Jupyter Notebook, and Python was the main programming language. Data extraction and preparation were exceptionally large parts of this process, and the following tools were used to help execute it.

**Tools and Libraries Used:**

* + Pandas: Used for data importation and manipulation, extraction, and preparation.
  + NumPy: Utilized for additional data handling and operations.
  + SciPy: Used for Pearson correlation to calculate correlations between variables.
  + Matplotlib: Used for data visualizations.

**Data Source Selection:**

* + CSV files were chosen from sources with the most complete and accurate databases.
    - Pro Football Focus, Pro-Football-Reference, Fantasy Pros

**Data Cleaning:**

* + Data cleaning was done within Jupyter Notebook.
    - Data was assessed by size, shape, and completeness.
    - Multiple features were removed for incomplete or irrelevant data.
    - Data frames were combined to create new data frames to be used for analysis.

**Outliers:**

* + Box-plots were created to analyze outliers and determine if they needed to be removed.

**Correlation Analysis:**

* + A person correlation analysis was performed to determine further features to use or remove.

Jupyter Notebook provided the tools and environment needed to perform these tasks. It allowed for the extraction and data preparation that would lead to the analysis that would be performed and provide the results needed to support the Hypothesis.

# E. Data Analysis Process

## E.1 Data Analysis Methods

To test the Hypothesis and determine if NFL wide receiver rookie production can forecast future NFL production by predicting future one-thousand-yard seasons. An ML model was created with results that accuracy can be measured in predicting wide receiver production.

**Analytical Methods Used:**

* + The Gradient Boosting Classifier was chosen to conduct the analysis because it can manage complex variables and relationships without the tendency to overfit. It additionally provides flexibility within the dataset.
  + Correlation analysis was also done to single out features that would be used to model the Gradient Boosting Classifier to return the best results.
  + Data Cleaning and Processing had to be accomplished before any analysis could be done. Because of the large number of imported files used, an overwhelming amount of cleaning and data frame combining had to be accomplished first. Because of this, a large amount of time was used for processing and combining data for further analysis.

## E.2 Advantages and Limitations of Tools and Techniques

**Advantages of Gradient Boosting Classifier:**

* Outstanding predictive accuracy, in addition to its ability to manage complex relationships. Superb flexibility and the ability to fine-tune the results to create an even better model.

**Limitations of Gradient Boosting Classifier:**

* Computationally, it uses a great deal of resources with larger datasets.
* Requires delicate tuning of hyperparameters.

**Advantages of Correlation analysis:**

* Cost-effect statistical analysis to determine the relationship between variables.

**Limitations of Correlation analysis:**

* It is impossible to examine within the analysis to see if other factors affect the variables causing the correlation.

**Advantages of Data Cleaning and Processing:**

* Cleaning and processing the data provides a more accurate and reliable data forecast.
* By processing data into new data frames, it allows for streamlined analysis.

**Limitations of Data Cleaning and Processing:**

* Time and effort are needed to clean data from multiple sources.
* Missing fields can cause the removal of otherwise helpful data.

## E.3 Application of Analytical Methods

Data Cleaning and Processing:

* Missing and unneeded values were removed. After data sets are cleaned and combined into data frames that could be used for additional analysis, input, and data processing are significant parts of the process.

Correlation Analysis:

* Determining the features that will be used in the model is an integral part of the analysis. This helps to remove unneeded and potentially adverse results.

Gradient Boosting Classifier:

* Implementation of the central part of the primary data analysis. Here is the core of our prediction model that can handle the different complex data sets and returns the results of features used to predict values based on inputs.

Validation and Testing:

* For training and testing, the eighty-twenty method is used to train and test the model. An accuracy score, mean squared error (MSE), and f-score are used to determine model viability—an accuracy score above 80%, an f-score of .80 or better, and an MSE below .20.

# F Data Analysis Results

## F.1 Statistical Significance

The statisticalsignificance of the model analysis was determined using the Gradient Boosting Classifier. For this test, the Null Hypothesis is, “There is no relation between NFL wide receiver’s rookie year metrics and future production in the NFL.”

The details used for the analysis of the model are:

* Model used: Gradient Booster Classifier
* Accuracy Score
* F-Score
* Mean Squared Error

Results:

* Accuracy Score: 90%
* Mean Squared Error 0.1
* F-Score: 0.9

Conclusion

* The results of this analysis show that this model can efficiently predict wide receivers' future value and, in doing so, reject the null Hypothesis. With an accuracy score of almost 89%, the model can forecast future NFL production within a reasonable amount.

## F.2 Practical Significance

The practical significance of the model applies to its real-world applications. When fantasy football players draft their lineups each year, they are guessing what second-year wide receiver is likely to break. With less than 20% of players ever having a one-thousand-yard season, having a model that can forecast these players with 90% accuracy is a massive advantage when drafting breakout players. These are usually late-round players who cost little draft capital, but if they break out, they can be league winners and completely change a player's fantasy football season. This model can be used as a tool, as a stand-alone tool, or with a larger package to help fantasy players.

## F.3 Overall Success

The success was to show if a model could be created that could tie rookie performance in the NFL of wide receivers to future NFL production, namely one-thousand-yard seasons, as a measurement of achievement. The result was achieved through data cleansing and statistical analysis through correlation and data modeling. This model proves to be accurate and predictive and can be a helpful tool that allows for a distinct advantage for fantasy football players.

# G. Conclusion

## G.1 Summary of Conclusions

This project aimed to create a predictive learning model using data from NFL rookie wide receiver seasons and forecasting future career success. This career success is to be measured in the form of one thousand receiving yard seasons. Using data from 2006 and comparing both rookie data to that of players' entire careers to create a database of players who have achieved this and use those statistics to help forecast up-and-coming players in the NFL.

## G.2 Effective Storytelling

Twenty-five visualizations were used in this project. Storytelling was supported by these visualizations that provided an explanation of data but also helped to determine what data to use and not to use. The following tools were used to represent the data graphically:

* Box-plot: Box-plot was used to look at outliers in detail to determine if some outliers would need to be removed. After further analysis, it was determined that these were true outliers and could not be removed from the data set. The outliers were the data needed to help create the model because these are the players the model was looking for.
* Scatter Matrix: This was initially used to generalize the data and look for correlation between the features. After examining several scatter matrices, some interesting correlations emerged that needed to be analyzed further.
* Heat map: This was used in conjunction with the Spearman-Spearman correlation to further investigate the correlation between features and one-thousand-yard seasons. After singling out the eleven features with the most significant correlation, I reran the heat map with only these eleven features to examine the correlations more closely.
* Scatter Plot: First, this was used to visualize each feature compared with players who had yet to achieve one-thousand-yard seasons. Blue dots represent players who have, and red dots represent players who have not. After looking at the plots, it became more apparent what features needed to be included in the model and what features were not.
* Scatter Plot: After the analysis was completed and the model was run against real-world data, a scatter plot was created for each feature to view the model’s prediction for rookie players in the 2023 season. Again, the blue dot represented players forecasted to have one-thousand-yard seasons, and the red dot represented those who did not.

These data visualizations were created using Matplotlib and Seaborn. Both tools helped us understand the data used to make critical decisions.

## G.3 Recommended Courses of Action

Based on the model forecast and accuracy of the model, two recommended courses of action can be pursued:

* **Integration of the model into real-world applications**. With an accuracy rating of 90%, the model is ready to be applied to real-world data and used each year. This tool, available to fantasy football players, can give them the edge they need to win their leagues each year.
* **Expand the model to additional years and positions**: Expand the model to include rookie and sophomore years to continue forecasting breakouts in their third year. Having an additional year of data can help find players who were injured or unable to receive playing time in their rookie season. Also, an adapted model could be used for additional positions, such as running backs and quarterbacks.

# H Panopto Presentation

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=c8b56737-d57b-4662-ba4f-b1ee0161fd5d

# References

The Sophomore Surge: Analyzing second-year wide receivers in fantasy football. https://football.pitcherlist.com/the-sophomore-surge-analyzing-second-year-wide-receivers-in-fantasy-football/

Museezer C. The Beast – Final WR1 Analytical Model: Dynasty Rookie Wide Receiver Rankings. Dynasty Football Factory. Published May 6, 2024. https://dynastyfootballfactory.com/the-beast-final-wr1-analytical-model-dynasty-rookie-wide-receiver-rankings/

Andrews B. Which metrics predict year 2 breakouts? – The Wrong Read, No. 74. RotoViz. Published December 31, 2022. https://www.rotoviz.com/2022/12/which-metrics-predict-year-2-breakouts/