Programming for Big Data

Big Data Analysis and Predictive Model for NFL Running Backs: Draft Likelihood and NFL Success

Ben Fitzgerald
s00244687@atu.ie
Link to GitHub Repository

Atlantic Technological University
Sligo, Ireland
May 2024

1 Problem Statement

The National Football League (NFL) Draft is an annual event where professional American football teams select eligible college football players to join their rosters. It consists of seven rounds, with each team selecting one player per round in a predetermined order based on their performance in the previous season (the worst team last season gets the first pick in the draft and so on). The NFL Draft is the primary means for teams to acquire talent, address roster needs, and build competitive teams. Making successful draft picks is essential for teams to maintain competitiveness and sustain long-term success in the highly competitive league, the NFL.

Teams often rely on subjective evaluations, limited statistics, and outdated metrics, leading to uncertainties in player selection and missed opportunities. The study conducted by Berri & Simmons (2011) highlights the challenges associated with assessing a quarterback's potential during the NFL Draft. After analysing nearly four decades of data, the researchers found no significant relationship between a quarterback's draft position and their subsequent performance in the NFL. This finding underscores the difficulty that teams face in accurately predicting a player's potential solely through traditional scouting methods. Even for the most important position of quarterback, traditional evaluation methods reveal their limitations, underscoring the necessity for advanced analytical techniques to better inform draft decisions.

The financial implications of drafting the wrong player can be substantial. For instance, in the 2024 NFL Draft, the number one pick is set to earn $38.5 million over four years, with $25 million guaranteed as a signing bonus. This hefty investment underscores the considerable risk involved in selecting the wrong player, as it could result in a significant loss of resources for the team. A striking example of such a misstep is evident in the 2010 NFL Draft where the St. Louis Rams selected Sam Bradford as the first overall pick. He secured a staggering $78 million contract, with $50 million guaranteed, before new regulations were introduced to mitigate such expenditures. Bradford won just 18 out of 49 games and was traded 4 years later, this exemplifies the costly repercussions of a misguided draft decision.

The NFL Combine serves as a crucial evaluation platform where college players showcase their athletic abilities to NFL scouts and team representatives through a variety of standardised tests. These tests include the 40-yard dash (speed and acceleration), bench press (upper body strength), and vertical jump (explosiveness and leaping ability). Additionally, players undergo drills to showcase their agility, flexibility, and positional skills. These tests provide NFL teams with quantifiable data to evaluate prospects' athletic abilities and potential for success at the professional level.

The project evaluates and predicts running backs eligible to be selected in the NFL Draft, leveraging machine learning techniques and big data analytics. By integrating college football statistics and NFL Combine data, the project aims to develop predictive models capable of assessing a running back's likelihood of being drafted and their potential success in the professional league. The focus on running backs allows for targeted analysis of position-specific attributes and metrics, providing valuable insights for NFL teams.

In a study analysing NFL Combine data, Sierer et al. (2008) found significant performance differences between drafted and non-drafted athletes across various drills. Drafted players, demonstrated superior performance in metrics like the 40-yard dash and vertical jump compared to non-drafted players. These findings suggest that combine performance metrics may predict a player's likelihood of being drafted. However, a study conducted by Kuzmits & Adams (2008) investigates the effectiveness of the NFL Combine in predicting the performance of players in the NFL. Examining data from a six-year period (1999-2004), the authors analyse the correlation between combine test results and NFL success across three positions, quarterback, running back, and wide receiver. The study finds no consistent statistical relationship between combine performance and professional football success for quarterbacks and wide receivers but does find strong correlations between the 40-yard dash time and NFL success for running backs. These studies highlight the importance of NFL Combine metrics for running backs. Integrating these findings, the project aims to merge Combine metrics with college football statistics, enabling a comprehensive assessment of running backs' potential and providing actionable insights for NFL teams during the draft process.

Traditional scouting methods have limitations, leading to uncertainties in player selection and missed opportunities. This project offers a distinctive approach by leveraging machine learning techniques and big data analytics to predict the likelihood of running backs being drafted and their potential success in the NFL. By integrating college football statistics and NFL Combine data in a big data framework, this project is unique in its approach in player evaluation, providing NFL teams with more accurate and data-driven insights during the draft process. Also, while most similar projects focus on quarterbacks, this study prioritises running backs, a position often overlooked in player evaluation.

The project will utilise Apache Spark for scalable data processing, enabling efficient handling of extensive datasets covering NFL Combine and Draft records from 2000 to 2023. This approach enables the exploration of diverse feature sets, enhancing prediction precision. Leveraging Apache Spark's distributed computing capabilities establishes a sturdy foundation for scalability and future expansion. This framework sets the stage for incorporating additional features and expanding predictive models to cover various NFL positions concurrently, allowing for a more comprehensive analysis. Furthermore, it opens avenues for including data from future years, enhancing predictive capabilities and ensuring the models remain relevant for ongoing assessments.

In summary, this project aims to enhance the player evaluation process in the NFL Draft by utilising machine learning techniques and big data analytics. The goal is to predict the likelihood of running backs being drafted and their potential success in the NFL. By integrating college football statistics and NFL Combine data and leveraging Apache Spark for scalable data processing, the project aims to develop predictive models that offer valuable insights for NFL teams, thereby improving their decision-making processes during the NFL draft.

2 Project Requirements

1. **Data Acquisition and Integration:**

    - Scrape NFL Combine and NFL Draft data from Pro Football Reference for each year from 2000 to 2023 (*Pro Football Reference* n.d.)
    - Extract player URLs from this data and use those URLs to scrape Pro Football Reference's sister site, Sports Reference to obtain all college statistics for Running Backs that competed in the NFL Combine or were drafted between 2000 and 2023 (*Sports Reference* n.d.).

2. **Data Cleaning and Transformation:**

    - Identify missing values, inconsistencies, and outliers in the dataset obtained from Pro Football Reference and Sports Reference.
    - Implement data validation checks to ensure the integrity and quality of the dataset after cleaning.
    - Convert the cleaned data into a structured format suitable for analysis and modelling, ensuring compatibility with machine learning algorithms and statistical techniques.
    - Normalise the features as necessary to ensure uniformity and comparability across different variables.

3. **Model Development and Training:**

    - Train the models using historical data on a player's combine and college stats and compare them to a player's draft status and also impact in the NFL.
    - Develop machine learning models to predict a running back's likelihood of being drafted and also their predicted impact in the NFL.

4. **Model Evaluation and Validation:**

    - Evaluate the performance of the trained models using appropriate metrics such as accuracy, and area under the curve (AUC).
    - Validate the Draft Likelihood model using the 2023 NFL Draft as a validation set to see how well it predicted running backs entering the 2023 NFL Draft and compare it to where they were selected.
    - Validate the Success at NFL Level model using data from running backs entering the 2023 NFL Draft and see their predicted level of output for their NFL Career versus how they did in their first year in the league.

5. **Model Deployment and Utilisation:**

    - Deploy the trained models to make predictions on 2024 NFL Combine data for running backs to see which running backs are most likely to be taken in this year's draft.

- Utilise the Success at NFL Level model to generate predictions on which running backs will make the biggest impact in the NFL which can be used by NFL teams for decision-making purposes.

6. **Scalability and Future Expansion:**

   - Design the project architecture with scalability in mind to accommodate a larger dataset encompassing future draft years and the ability to scale this to all positions not just running backs.
   - Plan for future expansions, such as incorporating additional features, refining models, and enhancing prediction accuracy for comprehensive player analysis.

3 Big Data Platform Selection

The selection of Apache Spark as the primary Big Data platform for this project is founded on its robust features tailored to meet the scalability, redundancy, and resiliency requirements inherent in NFL data analysis. Apache Spark's architecture aligns seamlessly with the project's objectives of managing extensive datasets consisting of NFL combine data, draft outcomes, player attributes, and college statistics. The project encompasses 24 years of NFL combine and draft data (from 2000 to 2023 inclusive), with plans for ongoing growth each year. As the dataset expands, Apache Spark's distributed computing paradigm proves crucial, enabling efficient parallel processing across a cluster of nodes to handle the increasing data volume. This scalability is particularly crucial given the project's ambition to cover all NFL positions and draft years, necessitating a robust infrastructure capable of accommodating future growth.

Furthermore, while college statistics have been scraped for running backs, the project aims to scale to incorporate statistics for players across all positions. The availability of college statistics is expanding, even for players from lesser-known colleges, suggesting rapid growth potential in the coming years. Apache Spark's horizontal scalability capabilities provide an optimal solution for accommodating this anticipated growth in data volume and complexity. By leveraging Spark's distributed computing framework, the project can effectively process and analyse a broader range of college statistics, enabling comprehensive player analysis and enhancing prediction accuracy for draft outcomes and player performance.

Utilising Apache Spark as the primary Big Data platform is a unique approach to NFL player evaluation and modelling. By harnessing the power of big data analytics, the project aims to provide NFL teams with more accurate and data-driven insights during the draft process. This approach allows for the integration of extensive datasets covering NFL combine data, draft outcomes, player attributes, and college statistics, enabling comprehensive player analysis and enhancing prediction accuracy for draft outcomes and player performance.

Apache Spark's fault tolerance mechanisms, such as Resilient Distributed Datasets (RDDs) and lineage information tracking, ensure continuous data processing by reconstructing lost partitions in the event of node failures or disruptions. In practical terms, Spark processes extensive NFL combine and draft datasets spanning multiple years, distributing tasks across a cluster. If a node fails during data processing, Spark maintains the analysis pipeline's integrity. During model training, Spark automatically redirects failed tasks to other nodes to ensure uninterrupted processing, with its MLlib and ML packages adeptly handling interruptions to safeguard prediction accuracy.

Apache Spark is the ideal choice for the NFL project due to its superior performance, ease of use, and diverse analytics capabilities. Unlike Hadoop, Spark's in-memory computing model enables efficient batch and real-time data processing, crucial for handling extensive NFL Combine, NFL Draft, and College Statistics datasets of thousands of players. Spark's unified analytics engine seamlessly integrates tasks like ETL, machine learning, and streaming analytics, perfectly aligning with the project.

4 Initial Design

The project aims to address the challenge of optimising NFL team drafting decisions by leveraging machine learning and big data analytics. Through careful analysis of historical NFL Draft and NFL Combine data, the goal is to develop predictive models capable of assessing a running back's likelihood of being drafted and their potential success in the professional league. The significance of this endeavour lies in the substantial financial implications associated with drafting decisions, leading to significant losses for teams. To achieve this objective, the project outlines several key requirements for the application. These include the systematic scraping and integration of NFL Combine and Draft data, thorough data cleaning and transformation processes to ensure dataset integrity, and the development of machine learning models to predict draft likelihood and success at the NFL level. By fulfilling these requirements, the project seeks to provide NFL teams with valuable insights to inform their decision-making processes during the draft.

In setting up the project, I used Python along with BeautifulSoup and Pandas libraries to scrape NFL combine and draft data from Pro Football Reference efficiently. I built functions to handle URL extraction and HTML table parsing, and I included caching to avoid making redundant web requests. This approach allowed me to gather data comprehensively from 2000 to 2023, covering over 7000 unique players from both the NFL draft and NFL Combine. I saved the NFL draft and NFL Combine datasets as CSV files and uploaded them to Google Drive for easy access and use in Apache Spark. I developed a separate script to gather players' college statistics by leveraging their college URLs extracted from the combine and draft datasets. To manage potential challenges with large-scale scraping, I introduced a time delay between requests, allowing for smoother data retrieval while respecting server limitations. Given the vast amount of unique players, I decided to narrow the focus to running backs to streamline the extraction process. These steps laid a strong groundwork for detailed analysis and modelling of running back prospects for NFL teams. I selected Apache Spark as the Big Data platform due to its scalability and fault tolerance features, which are crucial for handling the large datasets obtained from the NFL combine and draft. Spark's distributed computing model ensures efficient processing of the extensive player data, while its compatibility with Python enables seamless integration with our existing codebase. This choice ensures that our analysis remains robust and scalable, meeting the demands of NFL data processing and analysis. As the project advances, the dataset is anticipated to grow with more years of data, reflecting the continuous nature of NFL events. Additionally, given the various positions in the NFL, the project must account for a broad scope beyond running backs. This necessitates a sturdy big data infrastructure to handle the collection of college statistics for all players. This scalability need emphasises the project's dedication to addressing the evolving demands of NFL data analysis.

In addition to scraping NFL combine and draft data, significant focus is placed on cleaning and preprocessing the raw datasets obtained from web scraping. This involves identifying and handling missing values,

inconsistencies, and outliers to ensure data integrity and reliability. Data cleaning techniques such as column renaming, URL formatting, data type conversion, missing value handling, data imputation, filtering, joining dataframes, and vector assembling were utilised to prepare the data for further analysis and modelling. Furthermore, feature engineering is conducted to extract pertinent attributes such as combine metrics and college football statistics, which are meticulously selected and transformed to capture essential factors influencing a running back's draft prospects and NFL performance. Model selection is then undertaken to identify the optimal fit for predicting both NFL draft likelihood and success. To evaluate the draft likelihood model, I'll compare its predictions against the actual outcomes of the 2023 NFL Draft, assessing metrics like accuracy and precision. Specifically, I'll analyse the predicted draft likelihood scores against the players' actual draft positions to identify any discrepancies. To evaluate the success at the NFL level model, I'll assess its effectiveness by comparing predicted player impact in the NFL to their actual performance metrics, such as Career Approximate Value (CarAV) which measures a player's impact in the NFL throughout their career. Specifically, I'll identify the top-performing players in last year's NFL Draft based on their predicted CarAV values and their value in their first year in the NFL. To further validate the models, I will gather data on 2024 college prospects. I will test both the draft likelihood and success at the NFL level models on this upcoming dataset. This validation step ensures that the models remain robust and effective for current draft prospects, providing NFL teams with up-to-date insights for their decision-making processes.

5 Implementation Summary

5.1 Data Preparation and Consolidation

In setting up the data processing environment, a SparkSession was initiated to facilitate efficient data handling using Apache Spark. The NFL Combine and Draft datasets, obtained through web scraping, were loaded into Spark DataFrames for further processing. To streamline the datasets, irrelevant columns were removed to focus on essential player attributes. For clarity and distinction between the two datasets, common columns such as player name and position were renamed. Inconsistencies in URLs, with some starting with "http" and others with "https," were addressed to ensure uniformity. An outer merge operation was then performed on the datasets based on matching college or NFL stats URLs, preserving all players who either attended the Combine or were drafted.

To consolidate player information, the coalesce function was employed to merge corresponding columns from the Combine and Draft datasets. This ensured that every player had key attributes like name, position, and college, regardless of whether they were drafted or attended the Combine. Additionally, the height column was split into total inches to standardise the representation of player height. To provide further insights into the drafting process, a new "Drafted" column was introduced to indicate whether a player was drafted. Finally, data types were standardised by converting them to floats or integers as appropriate. These transformations were crucial for ensuring consistency and compatibility with subsequent analysis and modelling efforts, laying the groundwork for meaningful insights into player evaluation and drafting strategies.

Running back statistics were seamlessly integrated into the dataset, with meticulous data type conversions and checks for null or blank values ensuring consistency and completeness. Pro Day data supplemented the dataset, addressing gaps exacerbated by the COVID-19 pandemic (Pro Days are like mini Combines held by Colleges for only their players, these were popular for players during the pandemic). Redundant columns were removed, and missing values in specific metrics were rectified, enhancing dataset accuracy. Rows lacking crucial metrics, like the 40-yard dash time, were judiciously removed to maintain integrity. These steps optimised the dataset for analysis and modelling, providing insights into player evaluation and drafting strategies. Additionally, the unique conference names that players participated in during college were extracted and categorised into Power 5 and non-Power 5 conferences.

5.2 Draft Likelihood Model

The correlation analysis revealed significant associations between several player attributes and the likelihood of being drafted. Notably, metrics such as the 40-yard dash time, average rushing yards per game, average scrimmage yards per game, and scrimmage touchdowns demonstrated positive correlations with draft likelihood. Conversely, the correlation between conference affiliation (Power 5 status) and draft likelihood was relatively weak.

```python
from pyspark.sql.functions import corr

# Compute correlation between each feature column and the target variable
correlation_results = {}
for column in ["forty", "G", "college_rush_att","college_rush_yds","college_rush_avg","college_rush_TD",
               "college_rec_rec","college_rec_yds","college_rec_avg","college_rec_TD","college_scrim_plays",
               "college_scrim_avg", "college_scrim_yds", "college_scrim_TD", "Power_5_Num"]:
    correlation_value = all_rb_df.corr(column, "Drafted")
    correlation_results[column] = correlation_value

# Print correlation results
for column, correlation_value in correlation_results.items():
    print(f"Correlation between '{column}' and 'Drafted': {correlation_value}")
```

```
Correlation between 'forty' and 'Drafted': -0.3723759023902145
Correlation between 'G' and 'Drafted': 0.017209719175177697
Correlation between 'college_rush_att' and 'Drafted': 0.1491916215983347
Correlation between 'college_rush_yds' and 'Drafted': 0.22352916309313797
Correlation between 'college_rush_avg' and 'Drafted': 0.18928895413518415
Correlation between 'college_rush_TD' and 'Drafted': 0.23551586729898139
Correlation between 'college_rec_rec' and 'Drafted': 0.042566781433639504
Correlation between 'college_rec_yds' and 'Drafted': 0.09931685737270803
Correlation between 'college_rec_avg' and 'Drafted': 0.1240170596290748
Correlation between 'college_rec_TD' and 'Drafted': 0.10612749208639789
Correlation between 'college_scrim_plays' and 'Drafted': 0.14524138857984778
Correlation between 'college_scrim_avg' and 'Drafted': 0.1676667191373251
Correlation between 'college_scrim_yds' and 'Drafted': 0.2239090554353721
Correlation between 'college_scrim_TD' and 'Drafted': 0.2422854645784708
Correlation between 'Power_5_Num' and 'Drafted': 0.05926833654065291
```

Figure 1: Likelihood Correlation

The logistic regression model utilises player-specific attributes like the 40-yard dash time, rushing attempts, yards, average, and scrimmage touchdowns to predict draft probability. Trained on labelled historical data, the model learns feature coefficients to estimate the likelihood of being drafted. It computes a probability score between 0 and 1 for each player, classifying them as drafted (1) or undrafted (0) based on a threshold. Evaluation via the Area Under ROC metric yielded strong performance, with validation and test scores of approximately 0.82 and 0.70, respectively. These metrics affirm the model's effectiveness in predicting draft likelihood and its value for player evaluation and drafting strategies.

The model's predictions offer valuable insights into draft likelihoods, with Keaton Mitchell being the notable outlier where he went undrafted even though according to the model he had a Drafted probability of 92.7%. Tank Bigsby on the other hand only had a drafted likelihood of 64.7% but was taken with pick 88. In the 2023 season, Keaton Mitchell went on to gain nearly 500 yards as a rookie whereas Tank Bigsby had less than 150. This isn't to say the model is perfect but there might have been an oversight by scouts which rated Mitchell too low and Bigsby too high for whatever reason. The model seemed in line with the draft predictions for the majority of the rest with no major outliers besides that.

```
Validation Area Under ROC: 0.81981981981982
Test Area Under ROC: 0.6984126984126985
```

```
# Show predictions for 2023 NFL Draft data
test_predictions.select("player", "Drafted", "prediction", "probability").show(truncate=False)
```

```
+------------------+-------+----------+----------------------------------------+
|player            |Drafted|prediction|probability                             |
+------------------+-------+----------+----------------------------------------+
|Tank Bigsby       |1      |1.0       |[0.3523098686064784,0.6476901313935216] |
|Chase Brown       |1      |1.0       |[0.18054121483856483,0.8194587851614352]|
|Zach Charbonnet   |1      |1.0       |[0.16923699923207136,0.8307630007679286]|
|Travis Dye        |0      |0.0       |[0.5775218429570995,0.4224781570429005] |
|Tiyon Evans       |0      |1.0       |[0.4305149125035872,0.5694850874964128] |
|Jahmyr Gibbs      |1      |1.0       |[0.13071199691223195,0.8692880030877681]|
|Evan Hull         |1      |1.0       |[0.23498105376842762,0.7650189462315724]|
|Roschon Johnson   |1      |1.0       |[0.3884866062201133,0.6115133937798867] |
|Kenny McIntosh    |1      |0.0       |[0.5029034682676263,0.49709653173237367]|
|Keaton Mitchell   |0      |1.0       |[0.07255755932954691,0.9274424406704531]|
|Camerun Peoples   |0      |1.0       |[0.23651085845384537,0.7634891415461547]|
|Deneric Prince    |0      |1.0       |[0.1609460132548103,0.8390539867451897] |
|Bijan Robinson    |1      |1.0       |[0.07758199776383749,0.9224180022361625]|
|Tyjae Spears      |1      |1.0       |[0.15687351153643916,0.8431264884635609]|
|Tavion Thomas     |0      |0.0       |[0.6466490023592654,0.35335099764073463]|
|SaRodorick Thompson|0     |0.0       |[0.5202225673812665,0.4797774326187335] |
+------------------+-------+----------+----------------------------------------+
```

Figure 2: ROC & 2023 NFL Draft Predictions

```
+------------------+----+----------------------------------------+
|player_name       |pick|probability                             |
+------------------+----+----------------------------------------+
|Keaton Mitchell   |NULL|[0.07255755932954691,0.9274424406704531]|
|Bijan Robinson    |8   |[0.07758199776383749,0.9224180022361625]|
|Jahmyr Gibbs      |12  |[0.13071199691223195,0.8692880030877681]|
|Tyjae Spears      |81  |[0.15687351153643916,0.8431264884635609]|
|Deneric Prince    |NULL|[0.1609460132548103,0.8390539867451897] |
|Zach Charbonnet   |52  |[0.16923699923207136,0.8307630007679286]|
|Chase Brown       |163 |[0.18054121483856483,0.8194587851614352]|
|Evan Hull         |176 |[0.23498105376842762,0.7650189462315724]|
|Camerun Peoples   |NULL|[0.23651085845384537,0.7634891415461547]|
|Tank Bigsby       |88  |[0.3523098686064784,0.6476901313935216] |
|Roschon Johnson   |115 |[0.3884866062201133,0.6115133937798867] |
|Tiyon Evans       |NULL|[0.4305149125035872,0.5694850874964128] |
|Kenny McIntosh    |237 |[0.5029034682676263,0.49709653173237367]|
|SaRodorick Thompson|NULL|[0.5202225673812665,0.4797774326187335] |
|Travis Dye        |NULL|[0.5775218429570995,0.4224781570429005] |
|Tavion Thomas     |NULL|[0.6466490023592654,0.35335099764073463]|
+------------------+----+----------------------------------------+
```

Figure 3: 2023 NFL Draft Probabilities and Outcomes

5.3 NFL Career Success Model

The career success model explores the correlation between player-specific attributes and the Career Approximate Value (CarAV), a metric indicating a player's overall career performance in the NFL. The correlation analysis shows a relatively strong positive correlation between all rushing and scrimmage stats but not as much from receiving stats. Similar to the Draft Likelihood model, playing in a Power 5 conference shows no correlation with career success in the NFL for running backs.

Predicting Career Success in the National Football League

```
from pyspark.sql.functions import corr

# Filter the DataFrame for data where year < 2018 because some running backs drafted in the
#last 5 years might not have accumulated enough carav yet for the ability they have
filtered_data = all_rb_df.filter(all_rb_df["year"] < 2018)

# Compute correlation between each feature column and the target variable
correlation_results = {}
for column in ["forty", "G", "college_rush_att","college_rush_yds","college_rush_avg","college_rush_TD",
               "college_rec_rec","college_rec_yds","college_rec_avg","college_rec_TD","college_scrim_plays",
               "college_scrim_avg", "college_scrim_yds", "college_scrim_TD", "Power_5_Num"]:
    correlation_value = filtered_data.corr(column, "carav")
    correlation_results[column] = correlation_value

# Print correlation results
for column, correlation_value in correlation_results.items():
    print(f"Correlation between '{column}' and 'Carav': {correlation_value}")
```

```
Correlation between 'forty' and 'Carav': -0.2810088045937276
Correlation between 'G' and 'Carav': -0.09092055831244437
Correlation between 'college_rush_att' and 'Carav': 0.20411541299235864
Correlation between 'college_rush_yds' and 'Carav': 0.3027694671185301
Correlation between 'college_rush_avg' and 'Carav': 0.23798473497412645
Correlation between 'college_rush_TD' and 'Carav': 0.27780909677060966
Correlation between 'college_rec_rec' and 'Carav': 0.08513333422879102
Correlation between 'college_rec_yds' and 'Carav': 0.14325903312917135
Correlation between 'college_rec_avg' and 'Carav': 0.2050265364611634
Correlation between 'college_rec_TD' and 'Carav': 0.15559636882327996
Correlation between 'college_scrim_plays' and 'Carav': 0.2025138988405899
Correlation between 'college_scrim_avg' and 'Carav': 0.19949186698983468
Correlation between 'college_scrim_yds' and 'Carav': 0.3038323683981451
Correlation between 'college_scrim_TD' and 'Carav': 0.29110513067197047
Correlation between 'Power_5_Num' and 'Carav': -0.012219370029319587
```

Figure 4: NFL Career Success Correlation

The linear regression model was built to predict NFL career approximate value (CarAV) for running backs. Trained on data up to the 2018 season and tested on the 2023 season, it used very similar metrics used in the draft likelihood model. The model's goal was to forecast the career success of running backs. After training, it was used to predict the career approximate values for players in the 2023 season, offering insights into their potential NFL careers.

The model predicts the same players at the top Robinson, Mitchel, and Gibbs as the draft likelihood model. Unfortunately, no current CarAV could be found for undrafted players so Keaton Mitchell does not have a

value for this. We can see Robinson and Gibbs are currently off to great starts in their careers after just one year accumulating 9 and 10 CarAV already so they were rightly seen as the top two running backs coming out of college and are on their way to fill their predicted potential.

```
# Order predicted players by highest prediction
predicted_players_ordered = predicted_players.orderBy(predicted_players.prediction.desc())

# Show the ordered predicted players
predicted_players_ordered.show()
```

```
+------------------+------------------+-----+
|            player|        prediction|carav|
+------------------+------------------+-----+
|     Bijan Robinson|30.104303523653826|    9|
|    Keaton Mitchell| 27.94328753944366| NULL|
|       Jahmyr Gibbs|26.745040541257808|   10|
|       Tyjae Spears| 24.31654558634017|    6|
|    Zach Charbonnet|23.703208358456635|    5|
|        Chase Brown| 22.26910728064884|    3|
|         Travis Dye|22.139310786694267| NULL|
|          Evan Hull|20.607034790308887|    0|
|    Camerun Peoples|17.272216034268638| NULL|
|        Tank Bigsby| 16.66056993964426|    1|
|     Deneric Prince|  16.4560733988843| NULL|
|     Kenny McIntosh|15.032549624515752|    0|
|    Roschon Johnson|14.432042474460332|    4|
|SaRodorick Thompson|12.436798619366783| NULL|
|        Tiyon Evans| 7.102339029453873| NULL|
|      Tavion Thomas| 6.867235730271787| NULL|
+------------------+------------------+-----+
```

Figure 5: Predicted NFL CarAV for 2023 Running Backs

5.4 Testing on 2024 College Prospects

Both models underwent testing on every running back who competed at the 2024 NFL Combine. There is no way to evaluate the results of these tests since the 2024 NFL Draft has not happened yet and the players have not accumulated any CarAV. The running backs entering the NFL in 2024 appear to have a lower overall quality compared to previous years, lacking a clear top star (Rang 2024).

Interestingly, the models present differing perspectives on the draft likelihood and potential career success of these prospects. For instance, while Jaylen Wright and Blake Corum are deemed most likely to be drafted, they rank lower in predicted CarAV compared to other prospects like Kimani Vidal. Vidal, despite having the highest predicted CarAV, is slightly behind Wright and Corum in draft likelihood according to the model. This difference highlights the importance of delving deeper into what factors influence draft picks compared to what truly determines success in the NFL. By understanding the metrics that scouts prioritise during player evaluations and contrasting them with those that correlate with NFL success, we can gain valuable insights. These insights can then help improve scouting strategies and player evaluation methods, ultimately making the drafting process more effective.

```
[71] from pyspark.sql.functions import col

     # Select desired columns and order by probability ascending
     ordered_predictions = predictions.select("player", "prediction", "probability") \
                                      .orderBy(col("probability").asc())

     # Show the ordered predictions
     ordered_predictions.show(truncate=False)
```

```
+---------------+----------+----------------------------------------+
|player         |prediction|probability                             |
+---------------+----------+----------------------------------------+
|Jaylen Wright  |1.0       |[0.12629462546423423,0.8737053745357658]|
|Blake Corum    |1.0       |[0.1340786393418934,0.8659213606581067] |
|Trey Benson    |1.0       |[0.13659251564568624,0.8634074843543138]|
|Isaac Guerendo |1.0       |[0.16939193726110463,0.8306080627388954]|
|Kimani Vidal   |1.0       |[0.17339639238329774,0.8266036076167023]|
|Keilan Robinson|1.0       |[0.18726126540762492,0.8127387345923751]|
|George Holani  |1.0       |[0.23317888239819712,0.7668211176018029]|
|Re'Mahn Davis  |1.0       |[0.2548769457399,0.7451230542601]       |
|Bucky Irving   |1.0       |[0.254925676670875,0.745074323329125]   |
|MarShawn Lloyd |1.0       |[0.27681576226508076,0.7231842377349192]|
|Tyrone Tracy Jr.|1.0      |[0.3119390887840139,0.6880609112159861] |
|Michael Wiley  |1.0       |[0.3417445246410042,0.6582554753589958] |
|Jawhar Jordan  |1.0       |[0.38932512069383585,0.6106748793061642]|
|Emani Bailey   |1.0       |[0.4066145378626602,0.5933854621373398] |
|Kendall Milton |1.0       |[0.4444170119502529,0.5555829880497472] |
|Audric Estimé  |1.0       |[0.4467487239841065,0.5532512760158935] |
|Cody Schrader  |0.0       |[0.5010435480949889,0.49895645190501114]|
|Dillon Johnson |0.0       |[0.60033538098549,0.39966461901451]     |
+---------------+----------+----------------------------------------+
```

Figure 6: Predicted Draft Likelihood for 2024 Running Backs

```
from pyspark.sql.functions import col

# Order the predictions by the predicted carav values in ascending order
ordered_predictions = new_predictions.orderBy(col('prediction').desc())

# Select the player and the predicted carav columns
ordered_predicted_players = ordered_predictions.select('player', 'prediction')

# Show the ordered players with the predicted carav values for the new DataFrame
ordered_predicted_players.show()
```

```
+---------------+-----------------+
|         player|       prediction|
+---------------+-----------------+
|   Kimani Vidal|24.154419223354864|
|   Bucky Irving|23.026807357213357|
|  George Holani|22.779697968290165|
|    Blake Corum|22.651178433564468|
|  Re'Mahn Davis| 20.79427836972117|
|Tyrone Tracy Jr.|20.049832235886157|
|  Jaylen Wright|19.751380736698167|
|  Michael Wiley|19.340228471190287|
|    Trey Benson| 18.75363483564702|
| Isaac Guerendo| 16.79794599202738|
| Dillon Johnson| 15.94496502121362|
|Keilan Robinson| 15.47364632362256|
|  Jawhar Jordan|15.457483317318946|
| MarShawn Lloyd|15.337823400302213|
|   Emani Bailey|14.557418798632355|
|  Audric Estimé|13.699186000494308|
|  Cody Schrader|11.829477688618596|
| Kendall Milton|10.869586769492741|
+---------------+-----------------+
```

Figure 7: Predicted NFL CarAV for 2024 Running Backs

6 Future Improvements

Looking ahead, there are key areas for advancing the project and maximising its impact in informing NFL team drafting decisions. Expanding the scope beyond running backs to include other positions is paramount. While the current focus is on evaluating running backs, incorporating data on quarterbacks, wide receivers, defensive players, and other positions will provide a more holistic view of draft prospects. By integrating position-specific performance metrics and college statistics for a diverse range of players, the predictive models can offer comprehensive insights to NFL teams across different positions.

In addition to NFL Combine data, integrating pro day statistics presents an opportunity to enhance the predictive models further. Pro days offer an alternative platform for college players to showcase their skills to NFL scouts and team representatives. By incorporating pro day metrics such as 40-yard dash times, shuttle run results, and position-specific drills, the models can capture a broader range of performance indicators and refine predictions for draft likelihood and NFL success.

Continued refinement and optimisation of the predictive models are essential for ensuring accuracy and reliability. Experimenting with different machine learning algorithms, fine-tuning model parameters, and exploring ensemble techniques will enhance prediction performance across various positions and datasets. Moreover, ongoing evaluation and validation using updated datasets, including recent draft outcomes and player performance metrics, will validate model effectiveness and inform further improvements.

Scalability and infrastructure optimisation remain critical considerations. As the dataset grows with additional years of NFL events and expanded coverage of player positions, leveraging cloud-based solutions and distributed computing frameworks will ensure efficient data processing and analysis. Implementing robust data pipelines that can handle large volumes of data from multiple sources, including NFL Combine, pro day events, and college statistics, is essential for scalability and real-time insights.

Also, exploring methods to gather Career Approximate Value (CarAV) for undrafted players is essential. Currently, it's challenging to obtain CarAV for undrafted players, as it's exclusively available within the NFL Draft tables. One possible solution could be to devise a unique CarAV formula. This seeks to enrich the thoroughness of player assessment and offer meaningful insights into the performance of undrafted athletes

Lastly, develop a user-friendly interface for NFL teams to interact with the predictive models. Creating a web-based dashboard or application that allows users to visualise model predictions, explore player profiles, and customise analysis based on position-specific criteria and preferences will facilitate informed decision-making during the draft process. By addressing these key areas of advancement, the project can continue to evolve and provide valuable support to NFL teams, ultimately improving drafting decisions and team performance in the league.

References

Berri, D. J. & Simmons, R. (2011), 'Catching a draft: On the process of selecting quarterbacks in the national football league amateur draft', *Journal of Productivity Analysis* **35**, 37–49.

Kuzmits, F. E. & Adams, A. J. (2008), 'The nfl combine: does it predict performance in the national football league?', *The Journal of strength & conditioning research* **22**(6), 1721–1727.

Pro Football Reference (n.d.).
URL: *https://www.pro-football-reference.com/*

Rang, R. (2024), '2024 nfl draft rb rankings: No clear stars, but deep top 10 prospects', *foxsports.com* . Updated Apr. 3, 2024, 2:47 p.m. ET.

Sierer, S. P., Battaglini, C. L., Mihalik, J. P., Shields, E. W. & Tomasini, N. T. (2008), 'The national football league combine: performance differences between drafted and nondrafted players entering the 2004 and 2005 drafts', *The Journal of Strength & Conditioning Research* **22**(1), 6–12.

Sports Reference (n.d.).
URL: *https://www.sports-reference.com/*