Prospect to Pro: Predicting NFL Success from College Football Performance

August 7th, 2024

—

Colin Montie
Carlos Garcia
Nicholas Davidson
Mattias Escudero
Nathan Wright

Overview

Our project aims to determine if there is a significant correlation between success in a college football career and subsequent success in an NFL career. To achieve this, we will collect and analyze numerous statistics and accolades from the college football careers of first and second-round draft picks. By comparing these data points to the players' achievements in the NFL, we hope to identify patterns and indicators that predict professional success.

Research Specifications

1. **Data Collection and Preparation:**

- Training Data: Gather comprehensive data on first and second-round picks from the 2013-2017 NFL drafts. This includes college and NFL statistics, awards, team success, injuries, and second contract offers.
- Testing Data: Collect data on first and second-round picks from the 2018-2020 NFL drafts to test our predictive models.
- Player Statistics: Compile detailed college and NFL stats for each player, such as game performance metrics, position-specific stats, and overall contributions.
- Awards and Accolades: Document awards and recognitions received during both college and NFL careers, such as All-American honors, Heisman Trophy, Pro Bowl selections, and All-Pro designations.
- Team Success: Evaluate the impact of team performance, including college bowl games and playoff appearances, as well as NFL playoff and Super Bowl participations.
- Injury Records: Track injury history and its impact on player performance and career longevity.
- Contract Data: Analyze second contract offers and other financial indicators to gauge professional success and stability.

2. **Data Analysis & Skill Development:**

- Utilize Python for data collection, cleaning, and analysis.

- Employ statistical methods and machine learning algorithms to identify and validate correlations between college performance and NFL success.
- Ensure that all team members gain hands-on experience with data programming and analysis techniques.
- Foster a collaborative learning environment where members can develop and refine their skills in Python.

3. **Industry Engagement:**

- Reach out to professionals in the football world, including scouts, coaches, and analysts, to gain insights and feedback.
- Collaborate with industry experts to refine our project development and enhance the accuracy and applicability of our findings.

Step-by-Step Research Process

Data Preparation:

- ➢ Located databases to acquire College Football/NFL Information
- ➢ Brainstormed which variables we would need from both College Football and NFL to aid us in our goal
- ➢ Web-scraped/pulled data from websites that matched with the variables we decided on, was done for the training data of the draft picks from the 1st & 2nd Rounds of the 2014-2020 NFL Drafts. Created a CSV with said information
- ➢ Defined “success” in a NFL Career as a career that consisted of a 1 First Team All-Pro Selection, or 37.5% Pro Bowl Selections in seasons played as a Starter in said career

Independent Variable Selection and Testing:

- ➢ Independent Variables
    - College Variables
        - Defensive
            - Solo Tackles_C: Number of unassisted tackles.
            - Assisted Tackles_C: Number of assisted tackles.
            - Total Tackles_C: Total number of tackles (solo + assisted).

- Tackles for Loss_C: Number of tackles resulting in a loss of yardage.
- Sacks_C: Number of quarterback sacks.
- Defensive Interceptions_C: Number of interceptions made.
- Passes Defended_C: Number of passes successfully defended.
- Fumble Recoveries_C: Number of fumbles recovered.
- Forced Fumbles_C: Number of fumbles forced.

■ Offensive

- Completions_C: Number of completed passes.
- Passing Attempts_C: Number of attempted passes.
- Completion Pct_C: Completion percentage of passes.
- Passing Yards_C: Total passing yards.
- Passing Yards per Attempt_C: Average passing yards per attempt.
- Adj Passing Yards per Attempt_C: Adjusted passing yards per attempt (factoring in touchdowns and interceptions).
- Passing Touchdowns_C: Number of passing touchdowns.
- Interceptions_C: Number of interceptions thrown.
- Passer Rating_C: Passer rating.
- Receptions_C: Number of receptions.
- Receiving Yards_C: Total receiving yards.
- Receiving Yards per Attempt_C: Average receiving yards per attempt.
- Receiving TDs_C: Number of receiving touchdowns.
- Rushing Attempts_C: Number of rushing attempts.
- Rushing Yards_C: Total rushing yards.
- Rushing Yards per Attempt_C: Average rushing yards per attempt.
- Rushing Touchdowns_C: Number of rushing touchdowns.

○ NFL Variables

■ Defensive

- Solo Tackles_NFL: Number of unassisted tackles.
- Assisted Tackles_NFL: Number of assisted tackles.
- Total Tackles_NFL: Total number of tackles (solo + assisted).
- Tackles for Loss_NFL: Number of tackles resulting in a loss of yardage.
- Sacks_NFL: Number of quarterback sacks.
- Defensive Interceptions_NFL: Number of interceptions made.
- Passes Defended_NFL: Number of passes successfully defended.

        - Fumble Recoveries_NFL: Number of fumbles recovered.
        - Forced Fumbles_NFL: Number of fumbles forced.
    - Offensive
        - Completions_NFL: Number of completed passes.
        - Passing Attempts_NFL: Number of attempted passes.
        - Completion Pct_NFL: Completion percentage of passes.
        - Passing Yards_NFL: Total passing yards.
        - Passing Yards per Attempt_NFL: Average passing yards per attempt.
        - Adj Passing Yards per Attempt_NFL: Adjusted passing yards per attempt (factoring in touchdowns and interceptions).
        - Passing Touchdowns_NFL: Number of passing touchdowns.
        - Interceptions_NFL: Number of interceptions thrown.
        - Passer Rating_NFL: Passer rating.
        - Rushing Attempts_NFL: Number of rushing attempts.
        - Rushing Yards_NFL: Total rushing yards.
        - Rushing Yards per Attempt_NFL: Average rushing yards per attempt.
        - Rushing TDs_NFL: Number of rushing touchdowns.
        - Receptions_NFL: Number of receptions.
        - Receiving Yards_NFL: Total receiving yards.
        - Receiving Yards per Attempt_NFL: Average receiving yards per attempt.
        - Receiving Touchdowns_NFL: Number of receiving touchdowns.
    - General Variables (College and NFL)
        - Draft Year: The year the player was drafted into the NFL.
        - Position: The playing position of the player.
        - Age: Age of the player at the time of draft.
        - Conference: The conference which the player was drafted from
        - All-Pros: Number of times selected as an All-Pro player.
        - Pro Bowls: Number of Pro Bowl selections.
        - NFL Success: Defined as having at least one First Team All-Pro selection or a Pro Bowl selection rate of 37.5% or higher in seasons played as a starter.
        - Career AV: Approximate value, a metric used to approximate a player's overall career value.

- Years as Starter (NFL): Number of years the player was a starter in the NFL.
- College Accolades: Awards and recognitions received during college career.
- Conference: The college athletic conference in which the player competed.
- National Championships: Number of national championships won during college career.
- Relative Athletic Score: A composite score evaluating a player's athleticism relative to their peers.

➢ Dependent Variable
  - NFL Success (Binary Variable)
    - 1 if Player has had 1 First Team All-Pro Selection or Pro Bowl Selections as starter in 37.5% of Career
    - 0 otherwise

Data Analysis - Model Testing, Training, Feature Selection:

➢ Model Training and Validation:
  - Split the data into training and testing sets (e.g., 80% training, 20% validation).
  - Train a range of models including Logistic Regression, Random Forest, [other correlation tests run]
  - Tune hyperparameters using grid search or random search to optimize model performance.

➢ Cross-Validation:
  - Implement k-fold cross-validation to assess model performance and prevent overfitting.

➢ Model Evaluation:
  - Evaluate model performance using metrics such as accuracy, precision, recall, F1 score, and ROC-AUC.
  - Analyze feature importance to understand which variables are most influential in predicting NFL success.
  - Create confusion matrices to visualize model performance, particularly true positive, false positive, true negative, and false negative rates.

➢ Model Selection:
  - Select the best-performing model based on evaluation metrics. Ensure that the model balances precision and recall for reliable predictions.

Prediction Model Results

Listed below are the features chosen via model selection along with their corresponding weight in the model for each position group.

- Quarterbacks
  - College Accolades -0.192287
  - Completions 0.008131
  - Passing Attempts 0.011942
  - Passing Yards per Attempt -0.010766
  - Passing Touchdowns 0.158126
  - Passer Rating -0.128005
  - Rushing Attempts 0.181881
  - Rushing Yards 0.765954
  - Rushing Yards per Attempt 0.128753
  - Rushing Touchdowns -0.262456
- Running Backs
  - College Accolades 0.33296
  - Receptions 0.535261
  - Receiving Yards 0.595535
  - Receiving Yards per Attempt 0.478276
  - Receiving TDs -0.007636
  - Rushing Attempts -0.221267
  - Rushing Yards 0.306797
  - Rushing Yards per Attempt -0.012136
  - Rushing Touchdowns -0.051761
- Wide Receivers
  - Age -1.054515
  - College Accolades 0.120254
  - Receptions -0.270274
  - Receiving Yards 0.027581
  - Receiving Yards per Attempt 0.365434
  - Receiving TDs 0.06835
  - Rushing Attempts -0.730706
  - Rushing Yards -0.401534
  - Rushing Yards per Attempt -0.602479

    - Conference (SEC) 0.947199
- Linebackers
    - Age -0.676045
    - College Accolades 1.044812
    - Solo Tackles 0.586625
    - Assisted Tackles -0.115703
    - Total Tackles 0.352792
    - Tackles for Loss 0.115439
    - Sacks 0.311067
    - Defensive Interceptions 0.445104
    - Passes Defended -0.19571
    - Forced Fumbles -0.18435
- Edge Rushers
    - College Accolades 0.685634
    - Solo Tackles -0.368864
    - Assisted Tackles -0.2076
    - Total Tackles -0.322227
    - Tackles for Loss 1.120693
    - Sacks -0.274965
    - Passes Defended -0.009542
    - Forced Fumbles 0.58936
- Cornerbacks
    - Age -0.211614
    - Solo Tackles 0.027618
    - Assisted Tackles -0.850231
    - Total Tackles -0.271188
    - Tackles for Loss 0.622164
    - Defensive Interceptions 0.212398
    - Passes Defended -0.186755
- Safeties
    - College Accolades -0.184117
    - Solo Tackles 0.241849
    - Assisted Tackles -0.331648
    - Total Tackles 0.041669
    - Tackles for Loss 0.502991
    - Sacks 0.051442
    - Defensive Interceptions -0.821128

- Passes Defended 0.667284
- Fumble Recoveries 0.158779
- Conference (SEC) 0.517931

- Defensive Tackles
  - Age -0.455977
  - College Accolades -0.137532
  - Solo Tackles -0.015585
  - Assisted Tackles 0.493228
  - Total Tackles 0.247453
  - Tackles for Loss 0.776103
  - Sacks 0.174414
  - Passes Defended 0.32125
  - Fumble Recoveries -0.991017
  - Forced Fumbles -0.434185

These models were used to predict whether or not the players in the test dataset would have successful NFL careers. The results of the classifications for each position group are shown below.

- Quarterbacks

```
Accuracy: 0.8333333333333334
ROC AUC Score: 0.5
Classification Report:
              precision    recall  f1-score   support

           0       0.83      1.00      0.91         5
           1       0.00      0.00      0.00         1

    accuracy                           0.83         6
   macro avg       0.42      0.50      0.45         6
weighted avg       0.69      0.83      0.76         6
```

- Running Backs

```
Accuracy: 0.6666666666666666
ROC AUC Score: 0.8
```

```
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.60      0.75         5
           1       0.33      1.00      0.50         1

    accuracy                           0.67         6
   macro avg       0.67      0.80      0.62         6
weighted avg       0.89      0.67      0.71         6
```

- Wide Receivers

```
Accuracy: 0.7692307692307693
ROC AUC Score: 0.6166666666666667
Classification Report:
              precision    recall  f1-score   support

           0       0.82      0.90      0.86        10
           1       0.50      0.33      0.40         3

    accuracy                           0.77        13
   macro avg       0.66      0.62      0.63        13
weighted avg       0.74      0.77      0.75        13
```

- Linebackers

```
Accuracy: 0.8333333333333334
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.83      0.91         6
           1       0.00      0.00      0.00         0

    accuracy                           0.83         6
```

macro avg	0.50	0.42	0.45	6
weighted avg	1.00	0.83	0.91	6

- Edge Rushers

```
Accuracy: 0.6666666666666666
ROC AUC Score: 0.5625
Classification Report:
```

	precision	recall	f1-score	support
0	0.70	0.88	0.78	8
1	0.50	0.25	0.33	4
accuracy			0.67	12
macro avg	0.60	0.56	0.56	12
weighted avg	0.63	0.67	0.63	12

- Cornerbacks

```
Accuracy: 0.7692307692307693
ROC AUC Score: 0.5
Classification Report:
```

	precision	recall	f1-score	support
0	0.77	1.00	0.87	10
1	0.00	0.00	0.00	3
accuracy			0.77	13
macro avg	0.38	0.50	0.43	13
weighted avg	0.59	0.77	0.67	13

- Safeties

```
Accuracy: 0.875
ROC AUC Score: 0.5
Classification Report:
```

	precision	recall	f1-score	support
0	0.88	1.00	0.93	7
1	0.00	0.00	0.00	1
accuracy			0.88	8
macro avg	0.44	0.50	0.47	8
weighted avg	0.77	0.88	0.82	8

- Defensive Tackles

```
Accuracy: 0.7777777777777778
ROC AUC Score: 0.5
Classification Report:
```

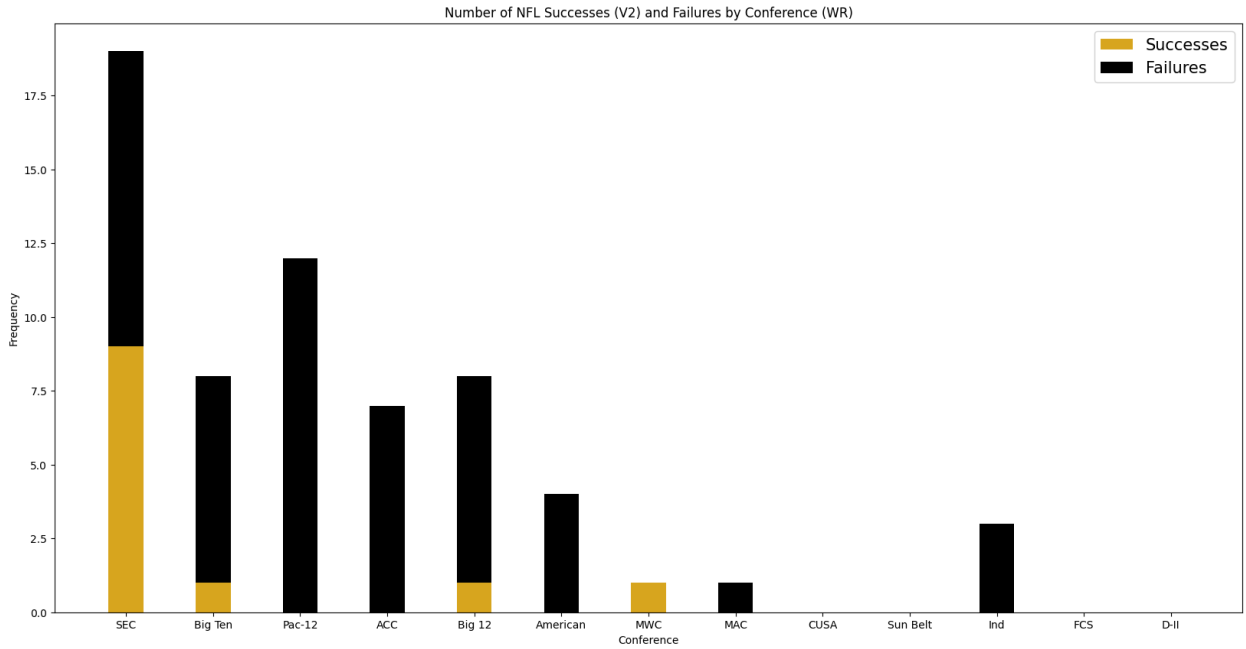
	precision	recall	f1-score	support
0	0.78	1.00	0.88	7
1	0.00	0.00	0.00	2
accuracy			0.78	9
macro avg	0.39	0.50	0.44	9
weighted avg	0.60	0.78	0.68	9

NOTE: Due to the small sample size of first and second round picks in the sample years, it is difficult to judge the true accuracy of the models. The accuracy scores shown above simply depict the percentage of test data points that were correctly classified by the models.

Probability Analysis Findings

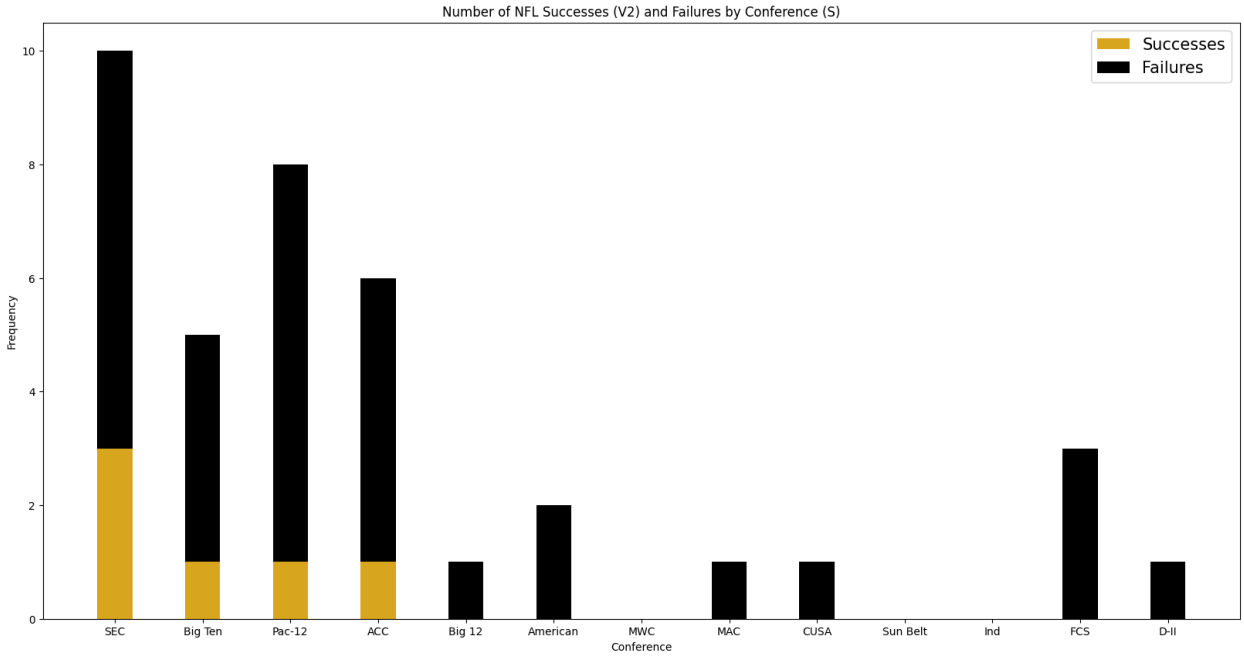
The following bar charts illustrate the distribution of NFL player outcomes from various college football conferences by position.

Wide Receivers (WR):



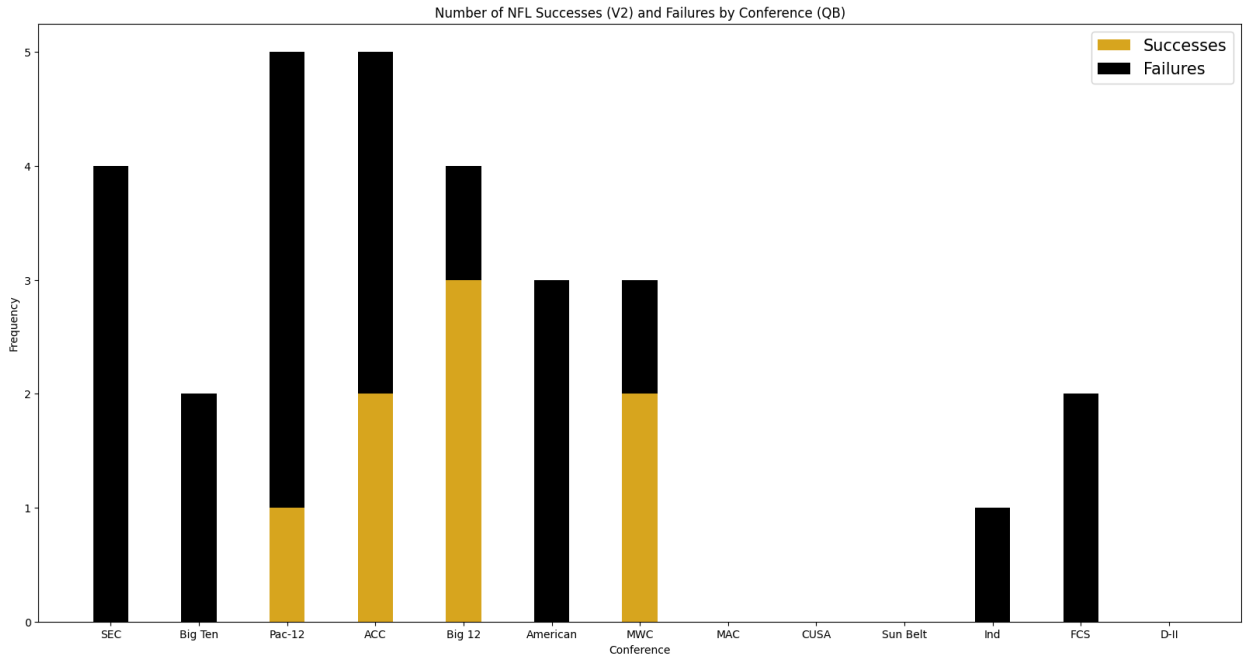
The graph reveals that the Southeastern Conference (SEC) is a major producer of NFL wide receivers, with approximately 7 successes and 18 failures. The Big Ten shows a different trend with about 2 successes and 7 failures. The Pac-12, ACC, and Big 12 also contribute significantly with more failures than successes, while conferences like the American and MWC have fewer but notable contributions, primarily showing failures.

Safeties (S):



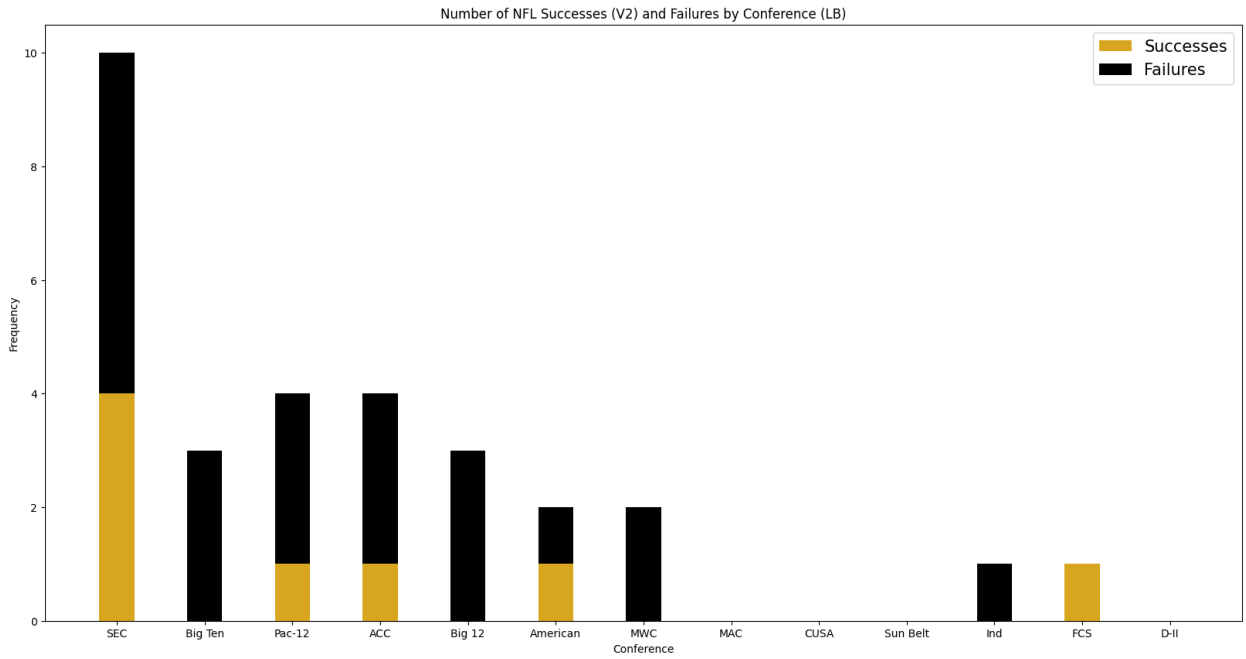
The chart indicates that the SEC leads in producing NFL safeties, with around 3 successes and 10 failures. The Big Ten and Pac-12 show a mix of results, each having approximately 1 success and several failures. Other conferences such as the ACC and American display similar patterns, highlighting challenges for players from these conferences in transitioning successfully to the NFL.

Quarterbacks (QB):



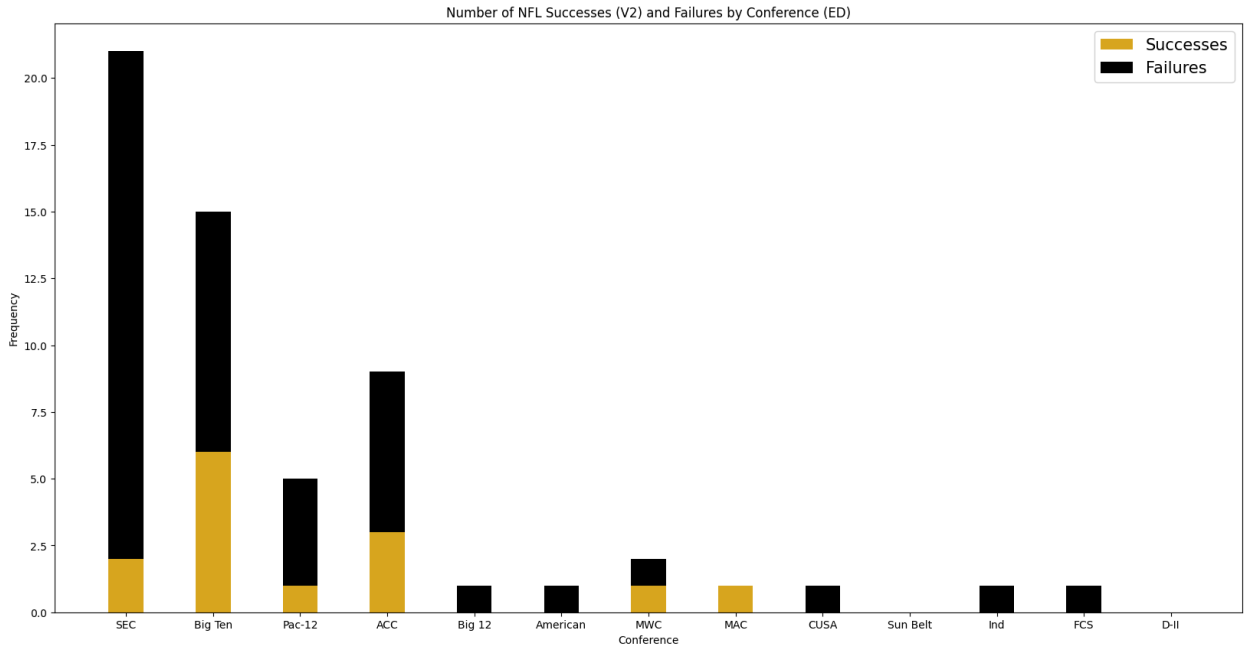
The graph shows the SEC as a dominant source of NFL quarterbacks, with about 2 successes and 5 failures. The Pac-12 and ACC also contribute notably, with each showing a balance of around 2 successes and 5 failures. Other conferences like the Big Ten and Big 12 have fewer contributions, with the Big Ten showing more failures and the Big 12 showing a balance between successes and failures.

Linebackers (LB):



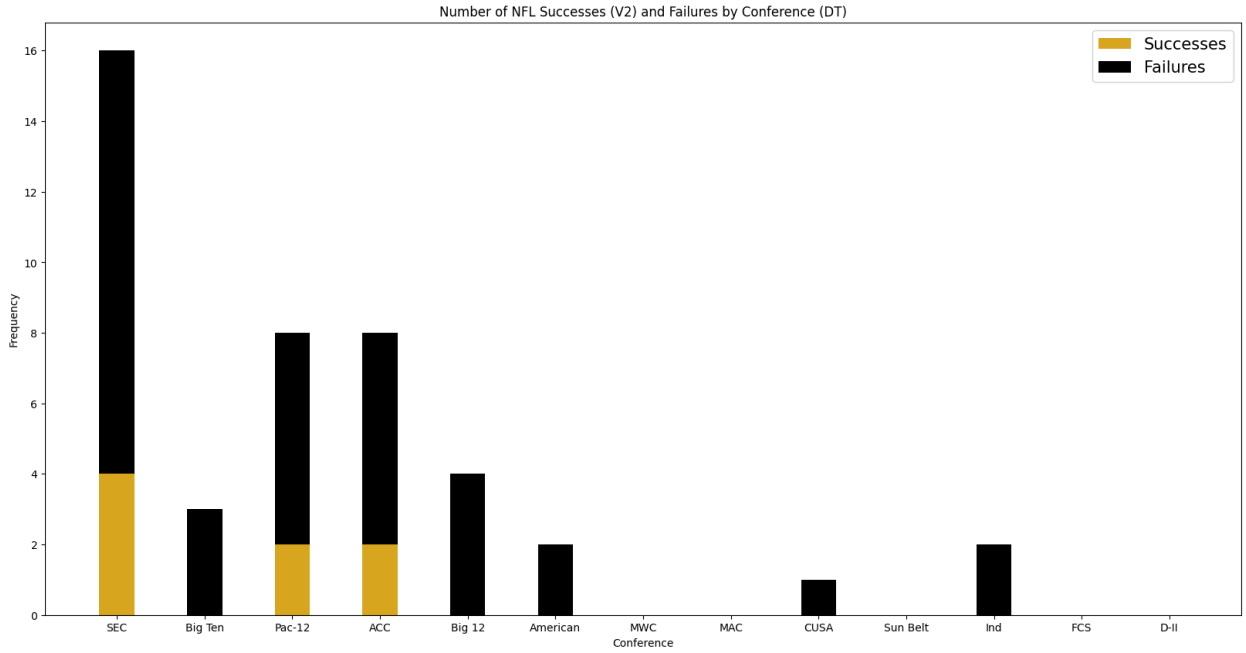
The chart illustrates that the SEC is the leading conference in producing NFL linebackers, with approximately 4 successes and 10 failures. The Big Ten and Pac-12 also contribute significantly, showing a combination of about 2-3 successes and 5-6 failures each. Other conferences like the ACC and Big 12 have fewer contributions, primarily showing more failures than successes.

Edge Defenders (ED):



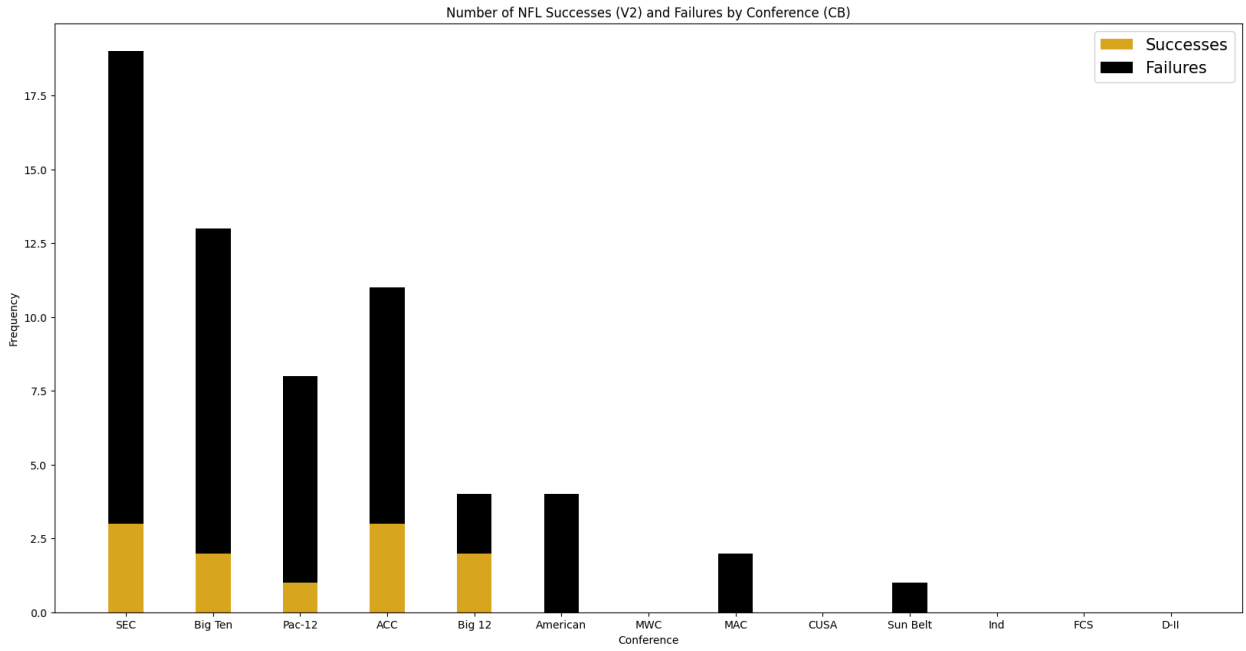
This graph highlights the SEC's dominance in producing NFL edge defenders, with about 4 successes and 20 failures. The Big Ten follows with around 3 successes and 12 failures. The Pac-12 and ACC contribute significantly as well, each showing more failures than successes, while other conferences like the Big 12 and American show fewer contributions, mostly failures.

Defensive Tackles (DT):



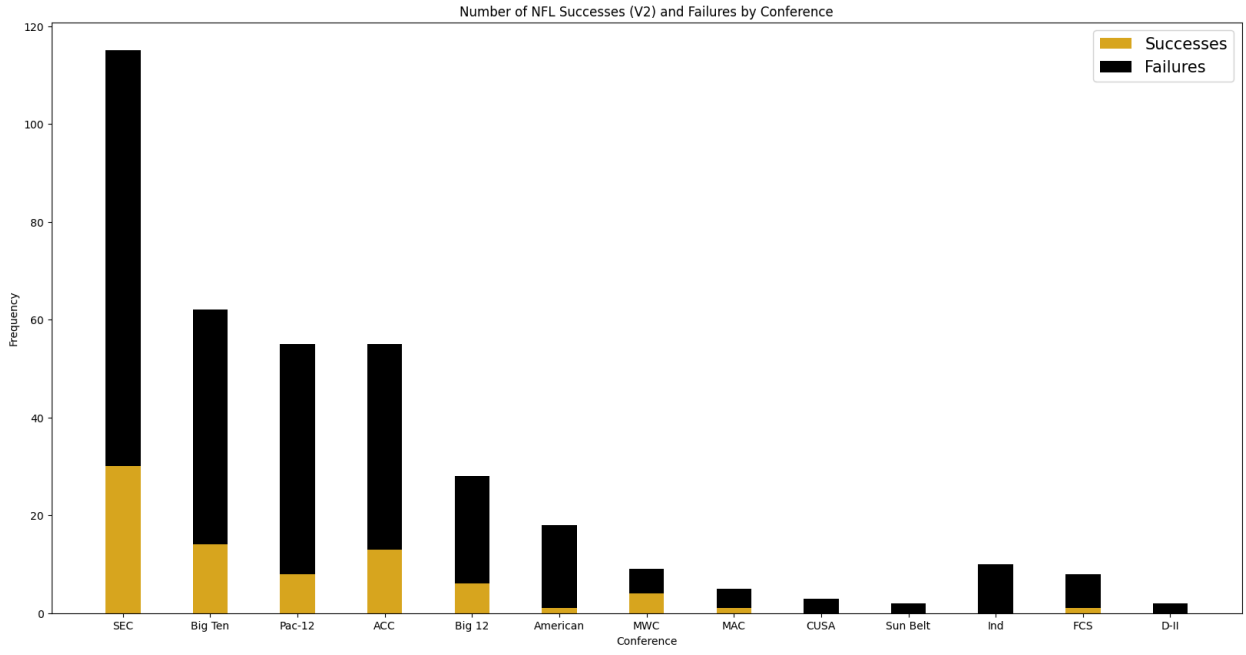
The chart reveals that the SEC is a significant producer of NFL defensive tackles, with around 3 successes and 16 failures. The Big Ten and Pac-12 also contribute notably, each showing a combination of 1-2 successes and 8-9 failures. Other conferences such as the ACC and Big 12 have fewer contributions, mainly showing more failures than successes.

Cornerbacks (CB):



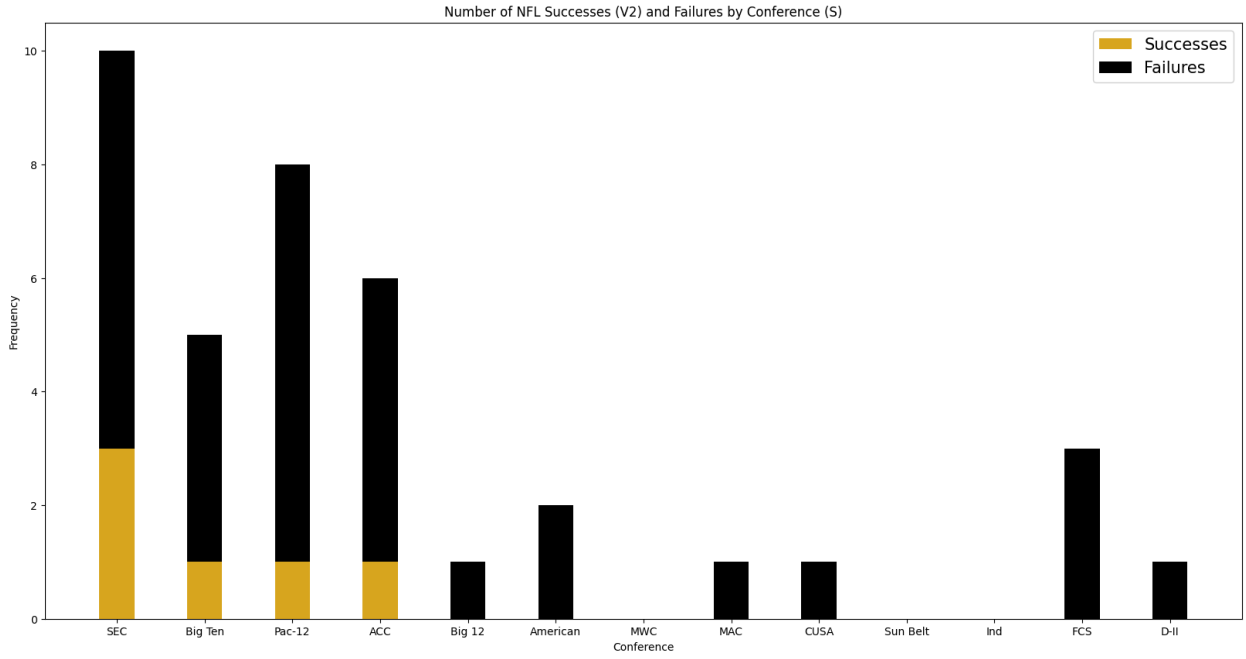
The graph shows that the SEC leads in producing NFL cornerbacks, with approximately 3 successes and 18 failures. The Big Ten follows with around 2 successes and 12 failures. The Pac-12 and ACC also contribute significantly, each showing a combination of about 1-2 successes and 8-10 failures, while other conferences like the Big 12 and American have fewer contributions, primarily showing more failures.

Overall Successes and Failures:



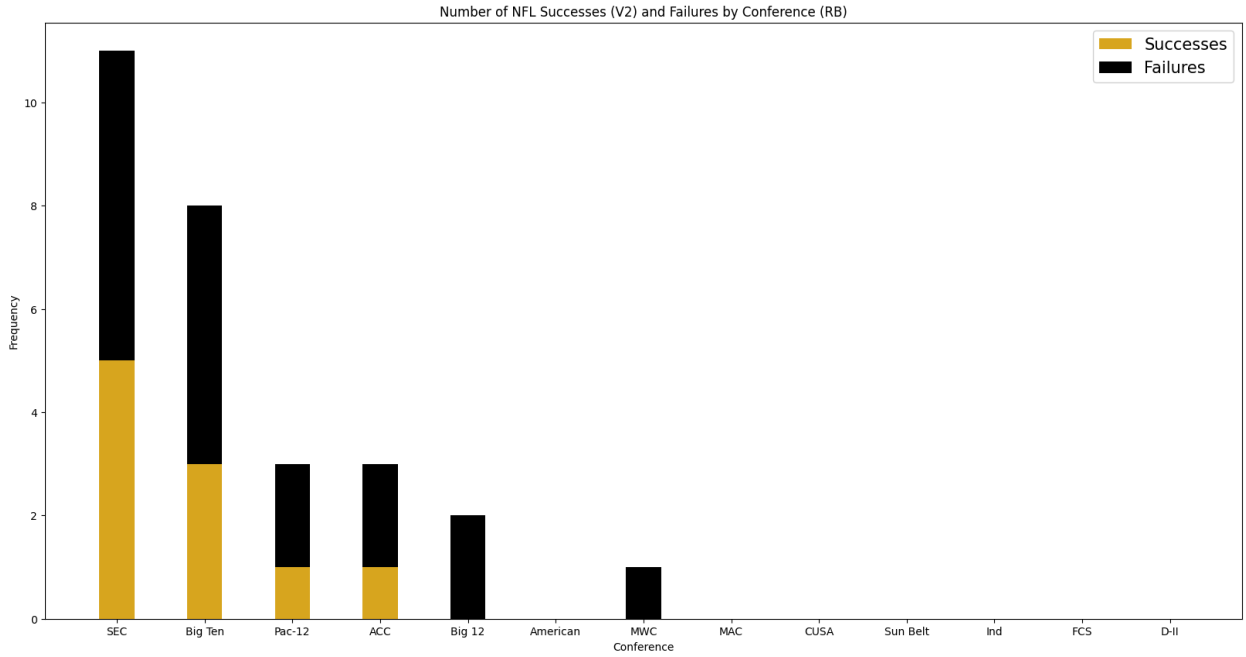
The comprehensive chart aggregates NFL career outcomes across all positions, revealing that the SEC is the dominant conference, with about 25 successes and 110 failures. The Big Ten also shows significant contributions, with around 10 successes and 60 failures. The Pac-12, ACC, and Big 12 display similar patterns, with more failures than successes, while conferences like the American, MWC, and FCS have fewer contributions, mainly showing failures, indicating the challenges players from these conferences face in achieving NFL success.

Safety (S):



The bar chart highlights that the Southeastern Conference (SEC) produces the most NFL safeties, with roughly 3 successes and 10 failures. The Big Ten, Pac-12, and ACC also show a higher number of failures compared to successes. Conferences like the American, MAC, and FCS have only failures recorded, while the MWC, CUSA, and Division II have very few or no data points.

Running Back (RB):



This bar chart reveals that the Southeastern Conference (SEC) dominates in producing NFL running backs, with about 5 successes and 10 failures. The Big Ten also contributes significantly, with around 4 successes and 8 failures. Other conferences, including the Pac-12, ACC, and Big 12, show a combination of successes and failures, whereas the American, MAC, and FCS mainly show failures, indicating difficulties for players from these conferences in achieving NFL success.

Conclusion

Our study aimed to identify which aspects of a college football player's career best correlate with success in the NFL, with the ultimate goal of developing a predictive model to assist NFL front offices in making better draft decisions. However, the main takeaway from our analysis was that there were few strong correlations between college statistics and professional success. This finding suggests that success in the NFL likely hinges on factors that are not currently quantified, such as mental aspects, scheme fit, and the quality of coaching at the next level.

Several limitations affected our study. We focused solely on first and second-round draft picks, which may have skewed our findings. Additionally, the dominance of certain players in specific positions limited the number of players deemed "successful" in our study, as seen with players like Travis Kelce at tight end and Tom Brady at quarterback. Moreover, many players' careers have not yet fully played out, making it difficult to measure their ultimate success accurately. One significant insight was the realization that individuals relying solely on publicly available data do not have a precise method for predicting NFL success.

Based on our findings, we recommend that teams avoid over-reliance on singular metrics and instead develop a composite score for prospect evaluation, incorporating a broader range of factors.

Future Possibilities

Looking ahead, several avenues for future research and model improvement present themselves:

1. **Data Expansion**: Expanding the dataset to include later-round picks and undrafted players could provide a more comprehensive view of the factors influencing NFL success. This broader dataset might reveal correlations that were not apparent with the more limited sample.
2. **Qualitative Factors**: Incorporating qualitative factors such as mental toughness, adaptability, and the quality of coaching received could enhance the model's predictive power. Future studies could explore methods to quantify these aspects or incorporate expert evaluations.
3. **Composite Score Development**: Developing a composite score that integrates various quantitative and qualitative metrics could provide a more holistic evaluation tool. This score could be tailored to include college performance, physical attributes measured at the combine, and other relevant factors.
4. **Longitudinal Studies**: Conducting longitudinal studies that track players' careers over a longer period would provide more data on players whose careers are still developing. This approach would allow for a better assessment of long-term success and the factors that contribute to it.

5. **Combine Data**: Testing and incorporating combine data could offer additional insights. Physical attributes measured at the combine, such as speed, agility, and strength, may provide valuable predictive information when combined with college performance data.
6. **Validation and Testing**: Future efforts should focus on validating the composite score or predictive model with new cohorts of players. Establishing benchmarks for success and continuously refining the model based on new data will be crucial for its practical application.
7. **Broader Impact**: Our findings highlight the importance of looking beyond readily available statistics and considering a wider array of factors in player evaluation. By influencing how NFL teams approach drafting and player development, this study can contribute to the broader field of sports analytics.

In conclusion, while our study did not find strong correlations between college statistics and NFL success, it underscores the complexity of predicting professional success and the need for a more comprehensive evaluation approach. By addressing the limitations and expanding the scope of future research, we can develop more effective tools for identifying and nurturing NFL talent.

Bibliography/Inspirational Resources

https://www.sports-reference.com/cfb/

https://www.pro-football-reference.com/

Draft Picks CSV (WIP)
https://colab.research.google.com/drive/1oVms1GmP4dBj4aA7hI7Cy0oYChCsez05?usp=sharing

https://www.sports-reference.com/blog/pfr-hof-monitor-methodology/

https://www.pff.com/