The Data Mine

ScoutSync: Redefining the Athletic Scouting Industry

Paul Amado | Dillon Barrow | Dominic Harbin | Sangeet Mohan | Nino Pizzo | Sam Shlau | Zarak Tareen | Drin Zhobe

OVERALL BACKGROUND

As it stands, the athletic scouting industry is extremely flawed and leaves great talent completely undiscovered. Our goal is to bridge this gap and ensure that *talent* gets opportunities, rather than just the players with the most hype or media attention. We have identified the most important physical and performance evaluation metrics in our focus sports, which include basketball, American football, and soccer. We then used these metrics to build models that could predict success based on real athleticism and production.

OVERALL FUTURE GOALS

Our overall goal is to broaden our findings to other sports such as hockey or tennis and for both men and women. On the football side, the long-term goal is to gather additional high school and college data to further train the models and refine them, such as in game statistics. With basketball, our goal is to expand our scores and research to high school and women's basketball, as data becomes available. And lastly, with soccer, our goal is to continue mapping our actual final production score value with pre-draft information that will be available to run comparisons with.

ACKNOWLEDGEMENTS

Darius Willis, Gizem O'Deorain - ScoutSync Project Mentors
- Provided professional guidance and support

Ram Muthurangan - Data Mine TA, Scrum Master
- Lead communicator and consult

FOOTBALL SUBTEAM

RESEARCH METHODOLOGY

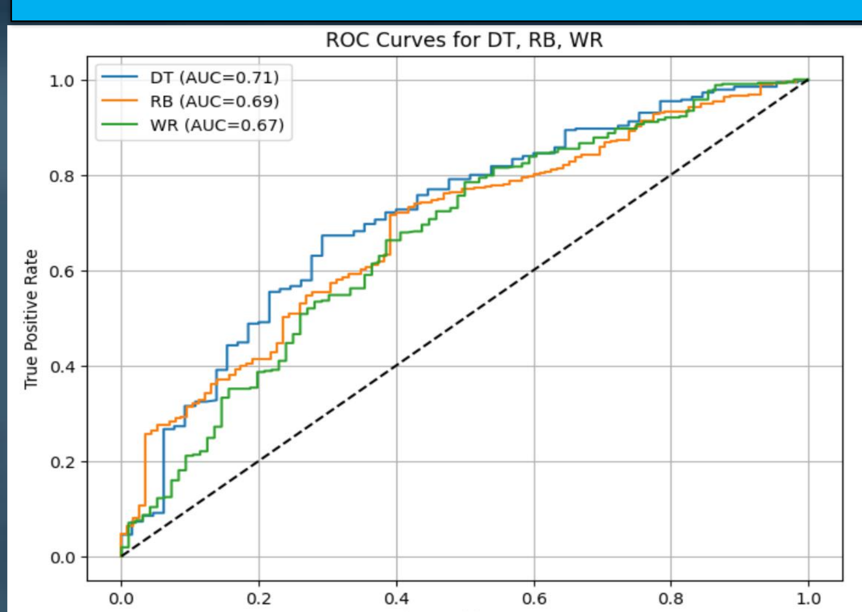
Collect Data
- Pro Football Reference Dataset Expansion (1987-2025)
- NFL Combine & Draft Data (1987-2025)
- Data Imputation (KNN)

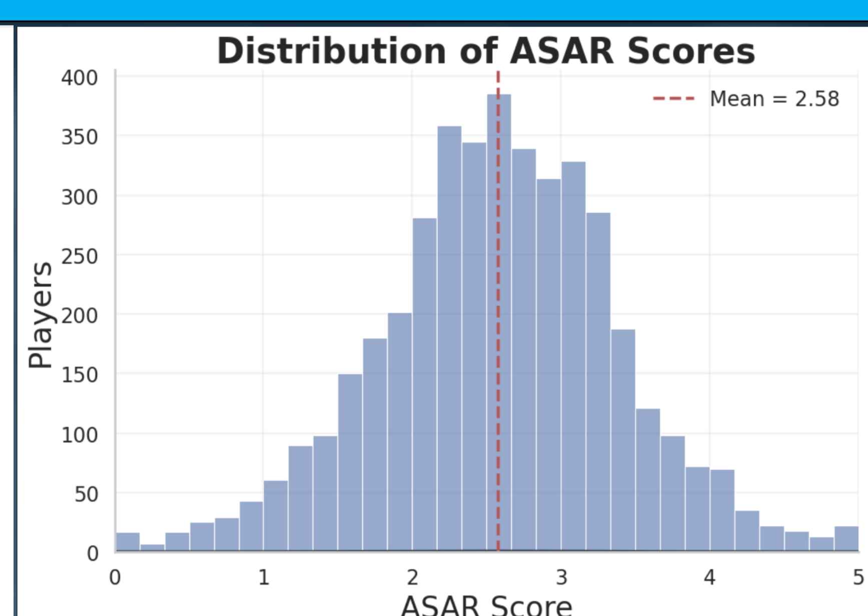
Create ASAR Score
- Used Speed, Agility, and Explosive Power Scores
- Logistic Classification (Target = Drafted)

Apply Formula
- Created new ASAR formula
- Applied to high school dataset
- Created pipeline to allow for user to input data

ASAR Score: Athletic Skills & Ability Rating (ASAR) score uses NFL Combine measurements to assign each player a 0-5 scaled comprehensive athleticism score.

ASAR ACCURACY AND DISTRIBUTION

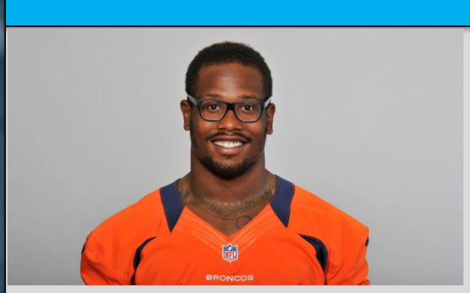




- Accuracy curves for DTs, RBs, WRs
- Model could fairly accurately predict drafted vs. undrafted based on metrics for these positions
- Model ranks a drafted DT over an undrafted DT 71% of the time

- Most college prospects show moderate athleticism
- Extreme scores (on either end) are relatively uncommon
- Rarity of high ASAR values effectively isolates true outliers and elite athletic profiles from the general population

CONCLUSIONS



Von Miller ASAR score: 4.98

While NFL Combine metrics are not definitive predictors of long-term success due to the sport's scheme-driven nature, they remain valuable for player evaluation. Our two-semester study found that athleticism correlates strongly with draft probability for specific positions. Consequently, we developed position-group formulas to calculate an Athletic Skill and Ability Rating (ASAR). This model is now used to identify high school prospects who might be overlooked by traditional recruiting.

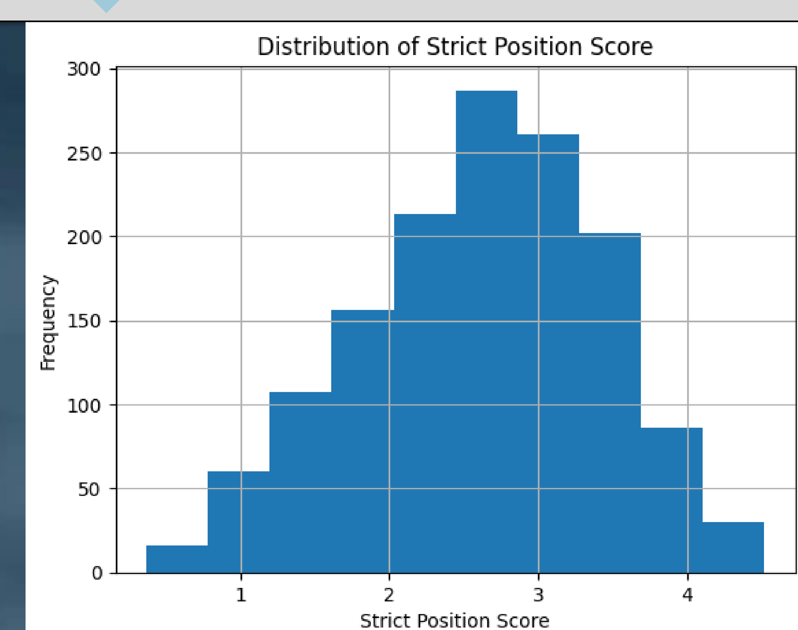
BASKETBALL SUBTEAM

RESEARCH METHODOLOGY

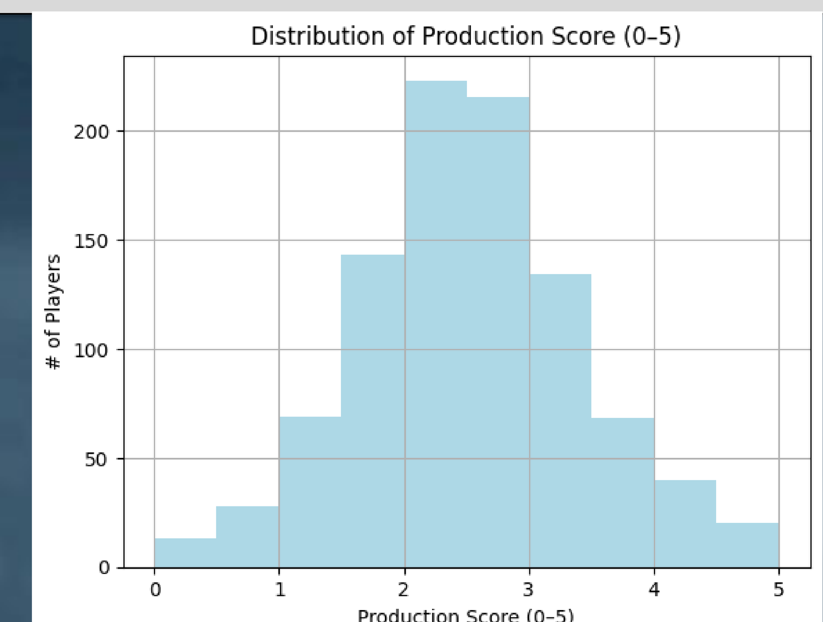
Collect Data
- NBA Combine Data 2000-2025
- NCAA MBB Statistics 2000-2025

Player Score
- Used Combine measurements to rank players on a 0 to 5 scale for size, speed, agility, and jumping metrics
- Created a final player score by combining each measurement score

Production Score
- Used NCAA game stats as a baseline (points, assists, and rebounds)
- Created production scores based on which stats are more important for each position



Player Score Model: A histogram of the distribution of player scores, from 0 to 5



Production Score Model: A histogram of the distribution of production scores, from 0 to 5

CONCLUSIONS

Highest Player Scores:
Guard: Scottie Lewis (4.51/5.00)
Wing: Hamidou Diallo (4.52/5.00)
Big: Kel'El Ware (4.49/5.00)

Highest Production Scores:
Guard: Ben Simmons (5.00/5.00)
Wing: Danny Granger (5.00/5.00)
Big: Anthony Davis (5.00/5.00)

Using the two scores we made, we were able to find evidence that these scores could be used in the future to train models, as well as predict where players would end up with certain metrics. With our player score, we are able to see how physical attributes, such as height and weight, combined with speed, jumping, and agility, can impact a player's ability to be drafted. With our production score, we can see how a player's NCAA statistics affect a player's draft stock, and we will hopefully be able to combine our player and production scores in the future.

SOCCER SUBTEAM

RESEARCH METHODOLOGY

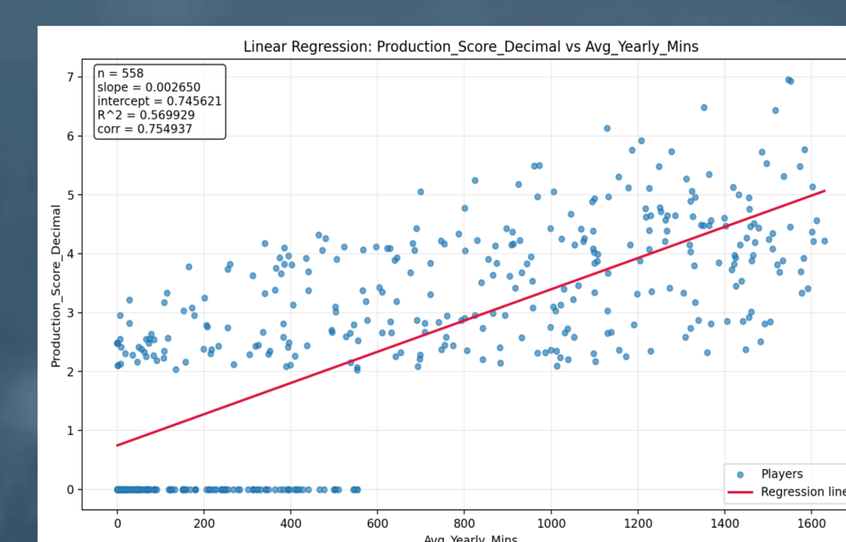
Collect Data
- 1998-2020 MLS "Superstars" Dataset
- NWSL Complete Dataset 2019-2025

Create Player Score
- Used Goals, Game Winning Goals, Assists, Shots on Goal, Shot Conversion, Amount of Yellow Cards, and Amount of Red Cards to determine a player's score

Apply Formula
- Run linear regression models on final production scores relative to career length and team wins
- Evaluate distribution of scores to ensure correlation with individual performance and team success

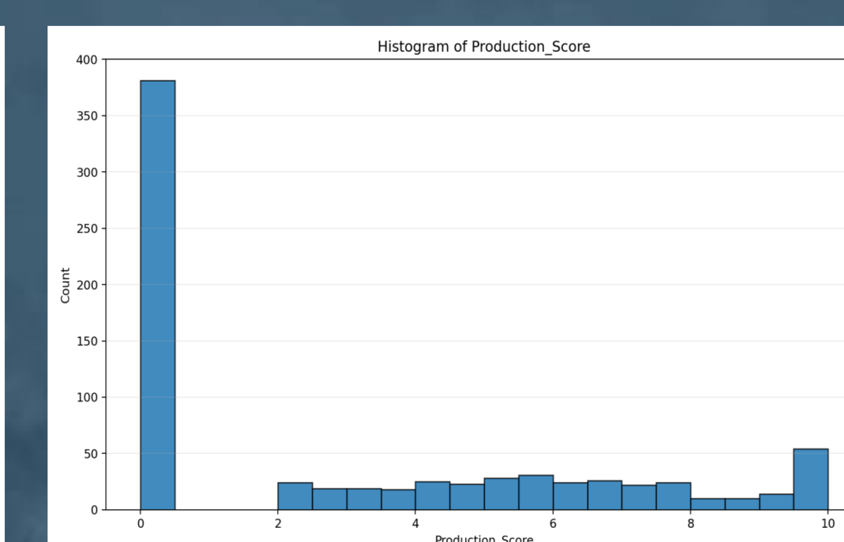
Linear Regression Model

(Plotting the production score calculated relative to minutes played)



Production Score Distribution

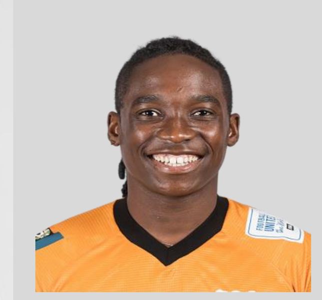
(Seeing overall how many players fell into each category of production)



PLAYERS THAT RECEIVED A 10.0 PRODUCTION SCORE

Alex Morgan

Mallory Swanson



Barbra Banda

Overall, I think that we can conclude that our creation and use of production scores can lead towards a more data-driven way to evaluate players. It's a scalable pipeline and can be tweaked for almost any given set of data we could be provided with. It produces objective metrics no matter what inputs it has, and that ultimately achieves the goal we set out at the beginning of the semester.

The Data Mine Corporate Partners Symposium