

Discipline in College Football

Discipline is a stat Pro Football Focus records to track a player's decision making during a game, penalizing poor on-field decision-making and execution. For my project, I am going to look at the relationship between discipline and success on the field in college football from 2022 to 2024. I will firstly establish the relationship between discipline and the other stats PFF tracks, as well as determine what types of players are the most or least disciplined (different positions, seniority, units, ...)

Broadly speaking I will use pandas and sql to get a sense of the data I'm working with, then move on to working with specific models (I anticipate using decision trees and regressions) to find the other stats/variables most correlated with discipline. All of my data will be gathered from PFF using their API, with each of my data sources being one of the years (2022, 2023, and 2024).

Index	Name	URL	Type	List of fields	Format	Collected?	Estimated Data Size
1	Player stats 2020	See below	API	See below	JSON	Yes, python	Tens of thousands of json entries
2	Player stats 2021		API		JSON	Yes, python	Same as above
3	Player stats 2022		API		JSON	Yes, python	Same as above
4	Player stats 2023		API		JSON	Yes, python	Same as above
5	Player stats 2024		API		JSON	Yes, python	Same as above

URLs:

https://api.profootballfocus.com/v1/grades/ncaa/2020/season_grade
https://api.profootballfocus.com/v1/grades/ncaa/2021/season_grade
https://api.profootballfocus.com/v1/grades/ncaa/2022/season_grade
https://api.profootballfocus.com/v1/grades/ncaa/2023/season_grade
https://api.profootballfocus.com/v1/grades/ncaa/2024/season_grade

Fields:

Index,player_id,season,position,unit,player,run_defense_snaps,pass_rush_snaps,coverage_snaps, total_snaps,run_block_snaps,receiving_snaps,pass_block_snaps,run_snaps,pass_snaps,run_defense, pass_rush,coverage,discipline,defense,run_block,receiving,pass_block,run,pass,offense,coverage_rank,offense_rank,defense_rank,run_defense_rank,pass_rush_rank,receiving_rank,pass_rank, run_block_rank,pass_block_rank,run_rank

Data Source Choice and Model Complexity Justification (Expanded)

I understand your interpretation of this being a single data source, and I choose to create a more complex model instead of choosing additional datasets. I have experience in modeling with my other courses, and I think I can make a high quality project even with just the one source.

Here is a more detailed outline of what I am planning to do with my project. Please let me know if there is anything that keeps it from being “complex enough” for the class, and please point me in the right direction in order to satisfy this threshold

Tree-Based Ensemble Models

I will employ Random Forests and Gradient Boosted Trees (e.g., XGBoost, LightGBM) to automatically capture non-linear interactions between variables without requiring manual feature engineering. These models are robust to multicollinearity and can handle complex hierarchical relationships between predictors, which makes them well-suited for modeling real-world educational performance data.

Regularized Regression Models

I will apply techniques like LASSO and Ridge regression for feature selection and interpretability, providing more insights to the ensemble methods. LASSO will allow me to zero out irrelevant features and highlight the most critical variables, while Ridge will help stabilize coefficients in the presence of multicollinearity.

Model Evaluation and Optimization

I will use cross-validation to ensure my models generalize well and to prevent overfitting. Additionally, I will perform hyperparameter tuning (e.g., grid search, randomized search, or Bayesian optimization) to systematically improve model performance and balance bias and variance.

Interpretability and Insight Extraction

I will conduct feature importance analysis, SHAP (SHapley Additive explanations) value calculations, and partial dependence plots to gain actionable insights on how different variables influence performance metrics.

By combining these techniques, I will ensure that even with a single dataset, my modeling process remains robust, interpretable, and sufficiently complex to meet or exceed project expectations. Please give additional feedback if needed. Thank you for your help.