Predictive Analytics for the 2024 ACC Championship Game:

A Machine Learning Approach

Amani Akkoub (ctf3un), Charli Ashby (arv2vp), Jonah Cicatko (zvh3zz),

Emily Hunter (brw6pe), Anne Kumashiro (yxt7ue), Jack Nickerson (rze7ud)

DS 3001 - Foundations of Machine Learning

Prof. Terence Johnson

December 14, 2024

Abstract/Executive Summary

The goal of this project is to predict the results of the college football 2024-2025 season, with a particular focus on the ACC conference. By leveraging historical data and statistical analysis, this study examines how trends in college football statistics contribute to the outcome of a game. This leads to the question: How do key season statistics influence the prediction of the ACC championship winner and score? This research concentrates on the Clemson and SMU football programs (the teams competing for the ACC title), utilizing game data from 2001-2024 to explore patterns that influence the result of the game. The primary aim is to apply these findings to predict the final results for the 2024-2025 ACC Championship Game. The study uses linear regression and decision trees models to uncover relationships within the data. Linear regression is applied to understand straightforward, linear connections between game factors and outcomes, while decision trees allow for the identification of more complex, nonlinear patterns. To prepare the datasets for analysis, preprocessing methods like one-hot encoding are employed to convert categorical variables into numerical formats. Principal Component Analysis (PCA) is also applied to reduce dimensionality and manage multicollinearity among features. The evaluation process relies on metrics like accuracy and the r^2 score, which measure the predictive performance of the models. Results from the analysis found that using PCA and Linear Regression to predict game scores was unreliable, with test r^2 values ranging from -0.44 to 0.30, indicating poor generalization and potential overfitting. While the models identified factors like passing touchdowns and kick return yards, the predictions lacked consistency, making it difficult to accurately forecast the ACC championship game's outcome. Despite challenges such as roster changes and evolving team compositions, the study not only identifies areas for improvement but

also attempts to establish a foundational approach for enhanced decision-making in college football analytics.

Introduction

College football presents a valuable opportunity to explore how data can explain and predict outcomes. Coaches and players change frequently, creating a variable environment for analysis. This study focuses on Clemson and SMU football programs to examine how game statistics influence results. The primary goal is to predict outcomes for the 2024-2025 ACC season and identify the most likely winner of the Championship game by answering the research question of: How do key season statistics (e.g., total yards, penalties, turnovers, and advanced statistics) influence the prediction of the ACC championship winner and score?

Sports analytics has advanced significantly in recent years, driven by improvements in data collection and analysis techniques. These advancements have enabled researchers to uncover meaningful patterns and relationships within complex datasets, providing actionable insights for coaches and analysts. By focusing on game factors such as offensive and defensive advanced statistics, this study aims to contribute to the growing field of predictive sports analytics. According to Connelly, a sports writer and analyst, the most important in-game statistics are explosiveness, efficiency, finishing drives, field position, and turnovers. In order to garner more accurate predictions, we attempted to correlate the statistics and outcomes of previous seasons.

This research makes three key contributions to the field. First, it examines the role of conferences in shaping game outcomes, highlighting that the opponent conference doesn’t always correlate to team success. This is important because, although the decision makers in NCAA football may not outright admit it, a conference’s strength affects their future playoff selection chances (Backus 1). Second, the study integrates advanced predictive models, such as linear regression and decision trees, with game-to-game updates to improve the adaptability and accuracy of predictions. These models are designed to capture both linear and nonlinear relationships within the data therefore, providing an understanding of the variables influencing outcomes. Third, it compares the performance of different predictive approaches. This offers insights into their respective strengths and limitations. By analyzing over two decades of data, the study not only explains historical trends but also generates predictions that can inform future strategies.

The significance of this study lies in its two-pronged focus on understanding historical patterns and predicting future outcomes. The findings are intended to help teams refine their strategies by identifying key performance indicators that contribute to success. The emphasis on actionable findings ensures that these insights are not just academic but have practical implications for improving team strategy and the sports betting industry.

One of the challenges addressed in this research is the ever-changing nature of college football. Teams undergo frequent shifts due to graduations, transfers, and injuries, which can complicate the process of drawing meaningful conclusions from historical data. This study acknowledges these limitations and incorporates methods to mitigate their impact, like that of updating datasets mid-season and incorporating new player statistics. By doing so, it offers a more accurate and adaptable framework for prediction.

The datasets used in this study span more than two decades, providing a rich foundation for analysis. By examining key metrics including scores, penalties, passing yards, and opponent performance, the study provides a full view of how different variables influence game outcomes. The combination of exploratory and predictive techniques ensures that the analysis is both thorough and actionable. These methods allow for the identification of patterns that may not be immediately apparent, offering a deeper understanding of the factors that drive success in collegiate football.

This study bridges the gap between historical analysis and predictive modeling, offering valuable insights for coaches, analysts, and sports enthusiasts. By utilizing the latest advancements in data analytics, it provides a framework for understanding and improving team performance. The findings not only highlight the importance of efficiency but also demonstrate the potential of data-driven approaches to revolutionize the way college football is analyzed and understood.

Data

The data used in this study were sourced from a publicly available API, (collegefootballdata.com) encompassing game statistics for Clemson and SMU from 2001-2024. We created datasets by calling the API's and separated them into groups for basic statistics and advanced statistics. Using the pandas package in Python, we saved data in a dataframe and converted it to a CSV that could be read later on. In the basic statistics' file we stored season, week, points scored, opponent name, as well as 32 additional statistics for the respective team and their opponent. These statistics included but were not limited to, total yards, turnovers, possession time, rushing and passing attempts, and fourth down conversions, offering a wide variety of team performance metrics across multiple seasons. The advanced statistics, composed of complex calculations collected by using more detailed player-tracking data, quantified havoc (plays that can change the momentum of a game for either team), explosiveness (how often a team generates big plays), points per opportunity, and success rate (the rate at which a play contributes to a first down). These statistics are much more nuanced in nature but because of their well researched formulas and the importance of these factors on the outcome of games, we felt them important to include in our advanced predictive analysis.

When we initially collected the data, we looked at each individual match-up and included the end of season statistics for either Clemson or SMU, next to the end-of-season statistics for their respective opponent as we assumed that those end-of-season stats would be a decent marker of how good a team was during the season we were analyzing. However, to accommodate for nuance and change throughout the season, we pivoted our approach by including the teams' statistics at the time of the match up. This would allow us to more accurately see the ups and downs in Clemson and SMU's performances throughout each season we drew data from.

Data preprocessing involved several steps to ensure readiness for analysis. This included resolving inconsistencies in variable formats, such as converting season identifiers to integers. Additional statistics, such as win-loss ratio variables were derived to enrich the data. We accounted for null variables that became null with the advent of new and more accurate statistics as College Football modernized. Categorical variables, such as conferences, were converted into numerical formats using one-hot encoding. PCA was applied to simplify data with overlapping information, addressing multicollinearity and ensuring statistical validity.

Methods

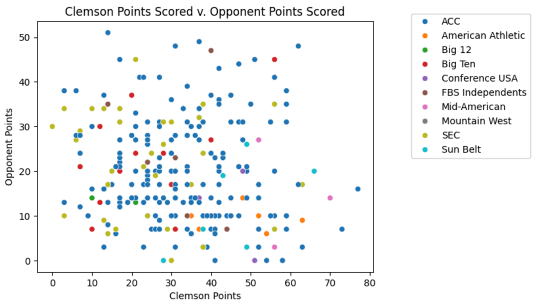
To achieve the research objectives, two primary predictive models were employed: linear regression and decision trees. Linear regression was effective in identifying straightforward relationships between game factors and outcomes. In contrast, decision trees excelled at capturing complex interactions, such as the combined effects of Clemson and SMU's as well as their opponents' statistics on results. Together, these models provided a comprehensive understanding of the variables influencing game outcomes.

To enhance predictive accuracy and manage high-dimensional season data, PCA was applied before employing linear regression. PCA helped simplify the data by combining related variables into fewer, more meaningful components while keeping most of the important information. This step made it easier for the model to work with the data while avoiding problems like overlapping effects between variables. After simplifying the data with PCA, we used linear regression to identify straightforward relationships between game statistics and results. We conducted PCA/Linear Regression on eight different dataframes. We permutated team (Clemson and SMU), data type (basic and advanced), and collection method (season total and weekly) in order to get the dataframes. In this way we were able to see which variables of data had the best accuracy using the resulting training and test r^2 values.

Decision trees were used to uncover more complex patterns in the data. Unlike linear regression, decision trees can figure out how multiple factors work together to affect game outcomes. By breaking the data into smaller groups based on specific conditions, decision trees created an easy-to-follow structure that revealed both simple and more complicated relationships in the data. We used decision trees to calculate relationships between team points and opponent points to the dataframe with the highest test r^2 value from linear regression. This was done as we inferred that the aforementioned correlation would be the most successful relationship, and thus would have the best chance of being accurate. By employing both PCA-enhanced linear regression and decision trees, the analysis benefited from a balance of simplicity and complexity.

Results

An exploratory analysis revealed that there was no correlation or clusters created when plotting Clemson’s points versus their opponents’ by conference (Fig. 1). A similar graph was also created for SMU (Fig. 2). The plot suggests that Clemson's scoring performance varies widely across games, with points ranging from low, even 0, to very high (up to 80). Opponent points also show variation but are generally lower than Clemson's, offering solid reasoning behind Clemson's dominance in many games. A similar result can be concluded from SMU, however, it can be seen that the green points, corresponding to the Big 12, are often located higher on the y-axis than the x-axis. This reveals that SMU tended to lose games against opponents within the Big 12.



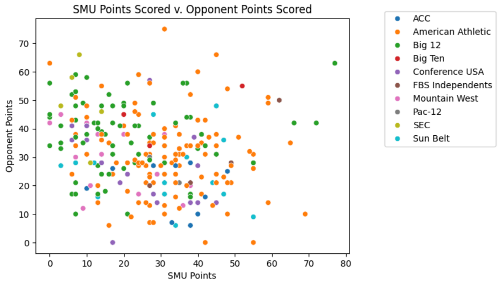
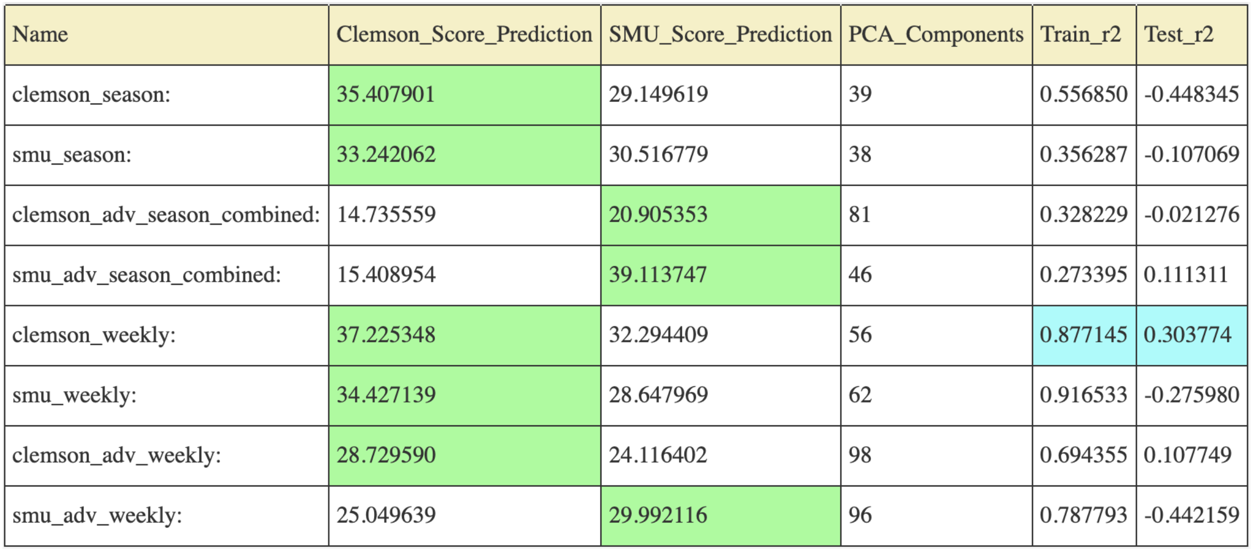


Fig. 1 - Clemson Points Scored v. Opponent Points Scored by Conference

Fig. 2 - SMU Points Scored v. Opponent Points Scored by Conference

The results of the PCA and Linear Regression showed that there was often low or no correlation between game statistics and score predictions. We created a table that displayed the names, score predictions, PCA components, training r^2 values and test r^2 values for each of the eight dataframes (Fig. 3). As shown, the r^2 of our prediction sets ranged from -0.44 to 0.30. This was significantly lower than our training sets. Therefore we cannot assume that this is representative of the situation and is likely overfit. Clemson's predicted score that resulted from these models, ranged from 15 points to 37 points with a mean of about 28.03, and a standard deviation of about 8.31. SMU's predicted scores, resulting from these models, ranged from 20 to 39. The scores had a mean of 29.34 and a standard deviation of 5.06. Although SMU has a higher mean, Clemson was predicted to win five out of eight times. This is likely due to a bias in the advanced statistics with season totals categories predicting Clemson to only score 15 points each time. Based on these contradictory markers, it would be highly difficult to predict a winner with our highly limited data. However, accounting for what our low outliers indicate could tip the scale in favor of Clemson.



Name	Clemson_Score_Prediction	SMU_Score_Prediction	PCA_Components	Train_r2	Test_r2
clemson_season:	35.407901	29.149619	39	0.556850	-0.448345
smu_season:	33.242062	30.516779	38	0.356287	-0.107069
clemson_adv_season_combined:	14.735559	20.905353	81	0.328229	-0.021276
smu_adv_season_combined:	15.408954	39.113747	46	0.273395	0.111311
clemson_weekly:	37.225348	32.294409	56	0.877145	0.303774
smu_weekly:	34.427139	28.647969	62	0.916533	-0.275980
clemson_adv_weekly:	28.729590	24.116402	98	0.694355	0.107749
smu_adv_weekly:	25.049639	29.992116	96	0.787793	-0.442159

Fig. 3 - Linear Regression Score Prediction Results

When creating the model for our trees to predict the most impactful statistics, we used the data from Clemson's basic data when cumulated weekly because it had a training R^2 value of 0.88 and a test r^2 value of 0.30. We used the CART algorithm to classify and create a regression. We found that for Clemson, the largest statistical factors that affected their score prediction were the season year, passing touchdowns, and kick return yards (Fig. 4). This didn't bring us a ton of clarity, but season could have been such a big factor due to covid as they were a higher-scoring team before 2020. We also created a tree that determined the greatest factors of opponents' scores (Fig. 5). These were opponent total yards, opponent third down conversions and kick return yards again. It was difficult to predict a numerical score using trees in this way. However, it did give us a baseline for our analysis.

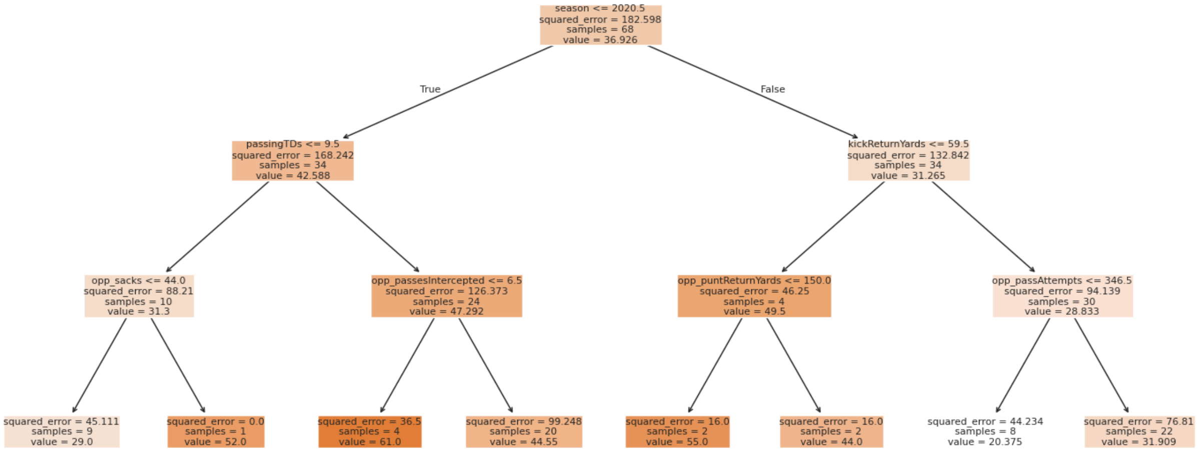


Fig. 4 - Clemson Decision Tree

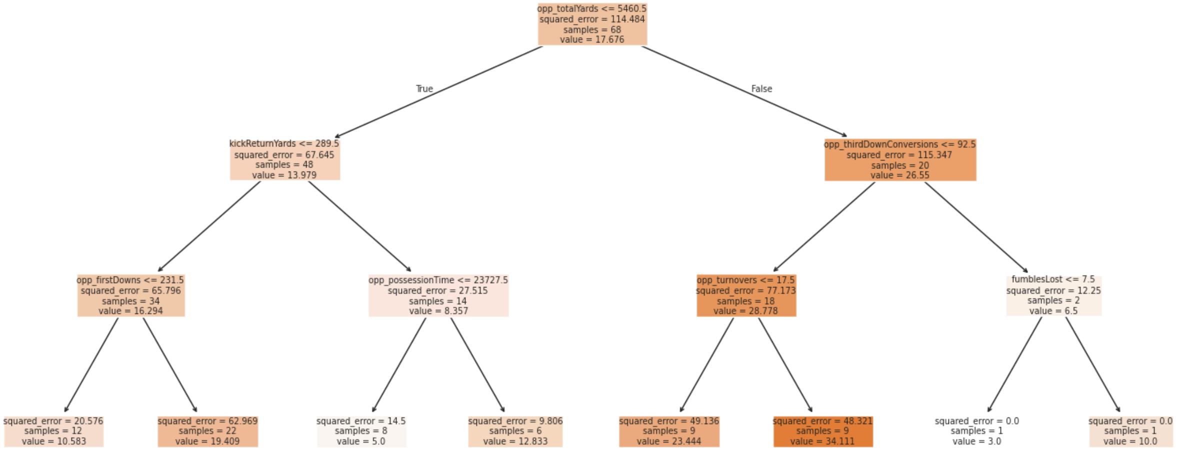


Fig. 5 - Opponent Decision Tree

The predictive models provided interesting insights. Linear regression effectively quantified the impact of continuous variables on game outcomes. The decision trees were able to identify complex patterns, albeit not the most helpful. Regularization techniques enhanced both models by addressing overlapping data and preventing overfitting. These findings demonstrate the importance of using multiple approaches to understand and predict game outcomes.

Conclusion

This study demonstrates the potential of data-driven approaches to enhance understanding and prediction of college football outcomes. By analyzing data from Clemson and

SMU, we identified basic statistics as being better at predicting game outcomes than advanced statistics. However, this did not have a high enough r^2 value to be a statistically significant correlation. Using both linear regression and decision trees provided insights into how different variables interact, while regularization ensured the reliability of the models.

The results of our analysis indicate that predicting game scores using PCA and Linear Regression models yielded inconclusive outcomes due to the low correlation observed between the game statistics and the predicted scores. The r^2 values for the test sets, which ranged from -0.44 to 0.30, suggest that the models did not generalize well to new data and may suffer from overfitting. Despite training r^2 values being significantly higher, the inconsistency in test performance showed a lack of predictive reliability. Additionally, while Clemson's predicted scores fluctuated between 15 and 37 points and SMU's between 20 and 39 points, the relatively close mean scores (28.03 for Clemson and 29.34 for SMU) make it incredibly difficult to determine a definitive winner. The models predicted Clemson to win five out of eight times, but this was potentially influenced by biases in the dataset, particularly from advanced season-total statistics. Moreover, CART analysis identified key factors like passing touchdowns, season year, and kick return yards for Clemson's score predictions, but these insights did not substantially clarify the score variations. Considering these findings, the limitations of our dataset and the inconsistencies in our predictive models highlights the difficulty within accurately projecting the outcome of the ACC championship game.

We determined that the study did not achieve its objectives, as it faced challenges such as team roster changes, missing data (past years where a statistic was not collected) and normal variance in sporting events which limited the predictive power of historical data. As the sports betting industry remains an unsolved multi-billion dollar industry, it was unlikely that a simple regression and decision tree would find meaningful results. However, we could continue future research that could address some of these concerns through the years by incorporating more detailed player statistics and expanding the analysis to include additional conferences could also improve the generalizability of the findings.

There are a few more possible challenges in this analysis, especially those stemming from conference play. Looking at the ACC over time, Stanford, California, and SMU are all new additions to the conference in 2024. Since the latter half of the season is filled with games against teams in conference, this limits our relevant data for these teams, as their results from previous years did not reflect their performance in this league, against these new teams. For example, SMU was previously a member of the American Athletic Conference, where they were a very competitive player in that league, but have never played against the Virginia Cavaliers, giving us very limited information in terms of their predicted performance against new opponents.

Something that comes up with narrowing our focus to the ACC is the new issue of analyzing out-of-conference performance. While the majority of varsity games occur within conference, there are a few out-of-conference games per season, varying drastically for each team. Some teams in the conference may have top-level out of conference competition, whereas others' may consist solely of what are called "tune-up games." This disparity in competition and lack of comparability is another thing we have to keep in mind when designing the model, since it is focused solely on a single conference.

An exploration into the change in the format would be valuable in the future because instead of the top 4 teams going to the College Football Playoffs like previous years, in the 2024-2025 season, the playoffs have expanded to 12 teams, made up of a combination of the 5 conference champions and the remaining 7 highest ranked teams. Attempting to predict the entirety results of the NCAA Division I (1,099 schools) would be an incredibly wide scope that would have a lot of beneficial insights.

This study offers a practical framework for using data analysis to understand and predict college football outcomes. By combining historical trends with predictive modeling, it provides insights for coaches, analysts, and fans. As sports analytics continues to evolve, adopting innovative methods and expanding data sources will be essential for unlocking new opportunities in competitive sports.

References

Backus, Will. (2024). 2024 College Football Conference Power Rankings: SEC starts season on top, but Big Ten isn’t far behind. CBSSports.com. https://www.cbssports.com/college-football/news/2024-college-football-conference-power-rankings-sec-starts-season-on-top-but-big-ten-isnt-far-behind/

CollegeFootballData.com. (2024). https://collegefootballdata.com/

Connelly, B. (2014, January). College football’s 5 most important stats. Football Study Hall. https://www.footballstudyhall.com/2014/1/24/5337968/college-football-five-factors