Modeling FBS Team Performance to Forecast Wins in Sports Analytics

LaNaysha Simms
Computer and Information Science
Responsibilities: Entire Project

December 2025

1 Introduction

Sports analytics is becoming increasingly important in the internet era, and with the landscape of college football, particularly within the Football Bowl Subdivision (FBS), changing every year, identifying patterns in how teams win and the factors that contribute to those victories is important for making informed decisions in team preparation throughout the season. In past research, Leung and Joseph [1] created their own algorithm that focused on four different factors derived from a game to predict the results of the bowl game and found great precision in predicting these results. Derived football attributes, such as rush-pass ratios, from the raw statistics of the game show great promise in predicting these outcomes, since there are many football statistics recorded that on their own that alone are not interesting, and these derived attributes fare well in common data mining algorithms used for prediction, such as linear regression and decision trees. In a study conducted in 2023, tree based models (decision trees, random forests) were found to show great accuracy in predicting win/loss and score differences, and Gifford and Bayrak [2] found around 80% success rates with these tests; the results in the study pertained to NFL games rather than NCAA games, but the meaning of the attributes provide the same insights with only semantic differences and are a great foundation for discovering patterns in college football outcomes. This report uses this background information and looks at the analysis of data with rolling averages (as well as semantic factors such as home/away rank, season week, etc.) in the season of a team home in an effort to predict not only the win/ loss of the team based on these factors, but also the difference of scores to see if averages throughout the season influence how a team performs in a game.

2 Methods

This project used Altair RapidMiner to facilitate the data preparation, mining, and results presented here. A diagram of the model is provided below in Figure 1, and starts with an encoding of a numerical value to nominal (season) to ensure the semantics of the values are used rather than the numerical values. The data allowed for empty values in the event that a team was unranked, and rankings are only awarded 1 through 25, so to treat all unranked teams as equal, a '26' ranking was given as a placeholder. The pipeline continues by replacing empty postseason values with a week '17', and to account for the newer era of playoffs that began in 2014 and to reduce the data to a relevant subset, the data from 2002-2013 was removed. To reduce the data further, two large conferences in the FBS, The Big Ten conference (B1G 10) and the South Eastern Conference (SEC) teams were only accounted for as the basis for the home team statistics (26 teams). In order to use rolling averages of the teams performance throughout the season, the data was first sorted by team and concatenated with year to serve as the variable by which to determine the averages per team/season for the weeks in their season.

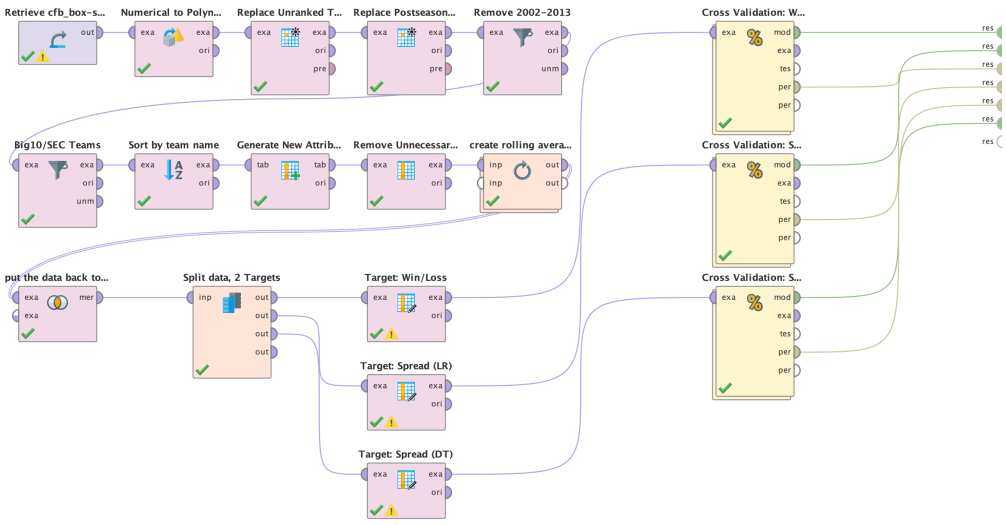

Figure 1: Visual Flow of Model in RapidMiner

New attributes were generated off of the raw attributes to get averages per game, such as spread(home score - away score, where positive score indicates a win and negative a loss), home_win, third down completion rate, etc. (a list of examples of these new attributes are provided below in Table 1, along with the formulas). Unnecessary attributes from the original data set such as time, attendance, and tv station were removed and then rolling averages were computed for the generated attributes; this data was used as the basis for the prediction models. In an effort to use the averages to predict both wins/losses and spread, the data was split and then used for three separate models that had cross validation with automatic sampling and 10 folds applied to them: one decision tree model to predict win/loss, one linear regression model to predict spread, and one decision tree model to predict score difference. On the original data, the decision tree model showed 81.78% accuracy on predicting win/loss of the home team, the linear regression model showed 14.006 RMSE on spread (+/- 0.350), and the second decision tree model showed a 16.700 RMSE on spread (+/- 0.814).

Derived Attribute	**Formula**
spread	score_home - score_away
home_win	if (spread $>$ 0, "Win", "Loss")
Home Third Down Completion Rate	if(third_down_att_home $>$ 0, third_down_comp_home/third_down_att_home, 0)
Home Fourth Down Completion Rate	if(fourth_down_att_home $>$ 0, fourth_down_comp_home/fourth_down_att_home, 0)
Home Pass Completion Rate	if(pass_att_home $>$ 0, pass_comp_home/pass_att_home, 0)
home_season	concat(home, "_", season)
Run-Pass Ratio	if(pass_att_home $>$ 0, rush_att_home/pass_att_home, 0)

Table 1: Derived Attributes and Corresponding Formulas

3 Experiments

The original data used was compiled by an independent user on the website Kaggle, found at this link:

College Football Team Stats Dataset (2002–2025)

and includes the records of college football games and their game statistics from 2002 - January 2025, with full game statistics available from 2004 - 2024 (previous games have empty entries for some of the stats

such as specific scores for a certain quarter). The games involved include records for every game associated with teams in the FBS (Football Bowl Subdivision, Division 1-A) in the NCAA (National Collegiate Athletic Association). The data was cleaned in the manner described in the methods section, and in an effort to see if the averages as a team performs throughout the year influence the performance in the later weeks, rolling averages of derived attributes from the raw attributes were computed. Since previous research pointed to using trees as a method of modeling data, decision trees were used to predict both win/loss and spread and the trees shown by RapidMiner provided much insight as to what was being considered in the prediction of these target variables. Linear regression was also used to see what factors had weight on predicting a numerical value, as well as to see if regression would perform better than decision trees in RMSE.

root_mean_squared_error

root_mean_squared_error: 14.006 +/- 0.350 (micro average: 14.010 +/- 0.000)

root_mean_squared_error

root_mean_squared_error: 16.700 +/- 0.814 (micro average: 16.717 +/- 0.000)

accuracy: 81.78% +/- 2.66% (micro average: 81.78%)

	true Win	true Loss	class precision
pred. Win	1566	284	84.65%
pred. Loss	195	584	74.97%
class recall	88.93%	67.28%	

Figure 2: Comparison of model performance; from top to bottom: Spread(Linear Regression), Spread(Decision Tree), Win/Loss.

With an 81.78% accuracy for preciting win/loss of a team, using the rolling averages did play a factor in determining game outcome, and in the breakdown of the tree (example path shown in Figure 3) showed that the rolling average spread, rank of the away team, and week of the game held the most weight in determining how the model made its predictions on win/loss, and with mispredictions it had trouble when it came to average points scored in quarters, as well as average pass completion rate in games. Since the accuracy rate rivaled what was deemed to be successful in previous research, further investigation was concluded, but there stands to be much more research that could be done to improve the accuracy of the model as well as looking at further reducing the parameters as has been done in previous work.

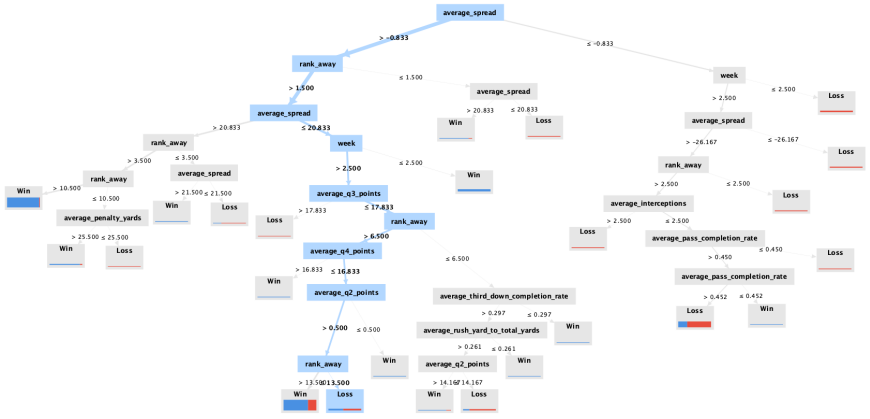

Figure 3: Example Path from the Decision Tree Model predicting Win/Loss

Looking at the prediction of spread with the decision tree, it had worse performance, with 16.700 RMSE

being predicted (or on average being around 17 points off of correctly predicting the spread). Rolling average spread held the most weight in this decision tree as well, but most of the other rolling averages and other attributes were more evenly distributed in determining the decision made by the model. In an attempt to get a smaller RMSE, linear regression was also used in determining the point spread, and provided a better RMSE with 14.006, but was not much better in determining the spread (ideally the model would have predicted within 5-10 points and would make the use of rolling averages hold more weight in determining the spread). A conclusion was reached in determining that although the rolling averages may be more relevant to the outcome of the game as it pertains to a home win, it does not align well with the prediction of spread for games; it is possible that looking at individual game averages rather than rolling game averages would provide better prediction of spread, and serves as a point of reference to look at for further work in this subject.

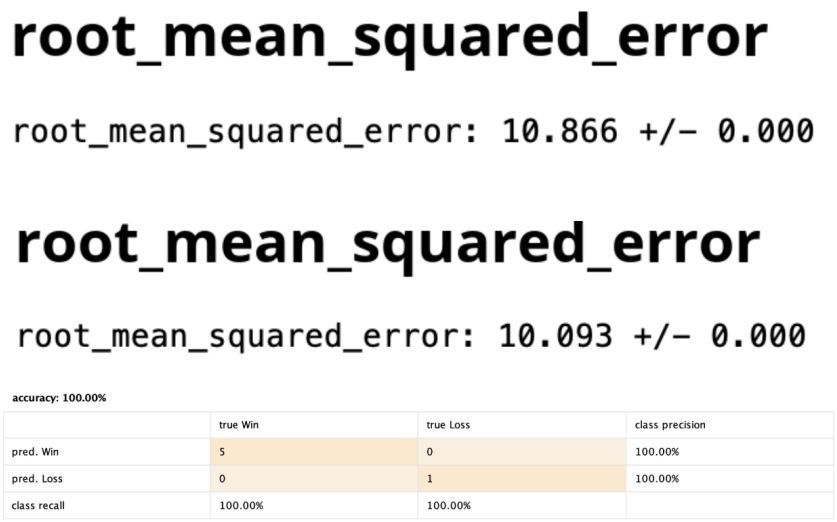

Figure 4: Model performance of Michigan 2025; from top to bottom: Spread(Linear Regression), Spread(Decision Tree), Win/Loss.

A final test was done on a self-compiled dataset that consisted of the home games from the University of Michigan football team's 2025 season, consisting of 6 games. As expected from the accuracy of the model, the model had a perfect prediction in determining win/loss; the tree did show that rolling spread played the biggest role in how the model made its decisions, and the dataset only had 6 items, so more new season data might provide a better picture of accuracy in the future. The RMSE values for the decision tree model showed better promise with the new data at 10.093, while the linear regression model gave 10.866; the difference between the two is very minuscule, but interesting to see the change in prediction, and this data fits closer to using rolling averages as a basis for spread prediction.

4 Conclusion

In summary, this project focused on using the concept of a home team's averages from game to game throughout the season with data mining to determine if it influences not only the outcome (win/loss) of a home game, but also the spread between the teams. Most of the work came from the preparation of the data into a format that matched what was intended to be tested, and involved learning much about the new software to me that is Altair RapidMiner as well as the mathematical concept of moving averages and implementing that in an efficient way.

One challenge I encountered was dealing with the curse of dimensionality and thinking that the original set of attributes that I had reduced to was small enough; I started trying to test with too many attributes at first and found that my models were taking forever to run, and was also working with over 10,000 instances

of data. I quickly realized that there was a reason that the previous research I had looked at used very small sets of games as basis for research, and did more work in understanding the domain of college football teams and statistics to lower my set to around 22 attributes from 58; I think that the set could still be reduced further after testing and am excited to continue working with this topic in the future now that I have a basis for further study. I also had trouble at first understanding how to implement the concept of rolling averages in the software since I knew the math it would entail from a theoretical standpoint but knew it would take far too long to implement myself. RapidMiner has great documentation, and I not only figured out how to use the software/functions to my advantage, but also get more comfortable with working with data mining concepts in general. As someone who is very interested in database management/analysis and is new to data mining, this project was very informative for my growth as a computer science student, and I look forward to continuing working with data mining and its concepts in the future.

References

[1] C. K. Leung and K. W. Joseph, "Sports data mining: Predicting results for the college football games," Procedia Computer Science, vol. 35, pp. 710–719, 2014.

[2] M. Gifford and T. Bayrak, "A predictive analytics model for forecasting outcomes in the National Football League games using decision tree and logistic regression," Decision Analytics Journal, vol. 8, p. 100296, 2023.