COSC 445: College Football Data Modeling

Kellen Leland

I. OBJECTIVE

The objective of this project is to collect various college football data, clean the data, and use it to create statistical models that will rank teams and predict weekly matchups. Prediction of weekly matchups will be separated into different categories: predicting which team will win the game, predicting the matchup against the spread, and predicting the matchup point total (over/under). Each of these models will have variations that can be compared and an analysis of each of the models' success will be completed at the end of the project.

II. MOTIVATION

I have always enjoyed watching and following college football. My father and I have a friendly competition each college football season. We each pick 10 games a week, and whomever has the highest percentage of accuracy at the end of the season is crowned as the winner.

In the past, I have always made my picks based on gut feelings and general knowledge of the various teams which has led to varying levels of success. I have recently become more interested in the data analysis side of these predictions as there are vast amounts of raw and compiled data available to the public for free.

I aim to improve the accuracy of my weekly picks by using digital archaeology techniques and data analytics in favor of my usual very subjective human analysis.

III. DATA DISCUSSION

College football is big business, and there is a vast amount of data available. Data will be pulled from publicly available APIs as well as scraped from various college football analysis websites and cleaned for use in the different statistical models. This data includes team offensive and defensive statistics, injury reports, forecasted weather for the game, previous matchup history, and much more. As previously mentioned, prediction will be split up into different

categories. Predicting which team will win the game is easy to understand, but I have used some terms which those uninterested or unfamiliar with college football may not understand.

Each week, various legal sports betting companies compile and release what they call the game ‘lines’. These lines include one that is called the ‘spread’ and another that is called the ‘total’. The ‘spread’ will determine a favorite - the team that is expected to win the game - and a number by which they believe they will win. For example, a spread could be described as “Tennessee -3”. This would mean that Tennessee is favored to win by a margin of 3 pts. Since Tennessee is favored in this example, the spread would be negative for them. The opposing team would get the same ‘spread’ but as a positive number since they are being considered the underdog. This means if I were to choose Tennessee -3, the Vols would have to win by more than 3 points for me to be correct. If I were to choose the opposing team or underdog with the positive spread, they would have to win or lose by less than 3 points and my prediction would be considered correct. The ‘total’ works very similarly. It will be set as a number of the predicted total points in the game scored by both teams combined. If I predict the under, I am predicting there will be less points than the total, vice versa for the over.

The lines are generally set by the various casinos and legal gambling companies so that 50% of people will bet on each side in order to minimize their losses and maximize their gains. These lines also eb and flow throughout the week as people place bets on either side, the companies will alter the lines to try to keep each side evenly taken. In order to simplify the process and models I create I will lock the lines on Tuesday of each week, and will not use the new values as they change throughout the week. This will also be another data point I can use in my favor in creating my predictive models. I will measure my success on each matchup against what the lines are when they are locked in on the Tuesday preceding the game.

IV. MEMBER RESPONSIBILITIES:

As I am working alone on this project, all responsibilities will fall to me. I will collect the data, clean the data, analyze the data, create the statistical models that predicts the matchups, and display the data in a way that is easy for the user to understand. I will also keep a record of each model's success in the

different categories and complete an analysis of what worked and what did not.

V. MILESTONES

Oct 15, 2021 - Complete collection of data

Oct 22, 2021 - Complete cleaning of data

Oct 29, 2021 - Complete predictive models

Nov 12, 2021 - Complete alterations/improvements to the models

Nov 22, 2021 - Complete rough draft of final report

Nov 26, 2021 - Add in analysis of each model's win/loss record

Nov 26, 2021 - Finalize report

VI. EXPECTED OUTCOME

I expect that by using a more objective and data focused approach in making my weekly college football matchup picks I can improve my overall accuracy percentage and have more consistent results when compared with my current subjective human analysis.