Springboard Introduction to Data Science

Capstone Milestone Report

This document will cover the section 6 requirement for discussing milestones for the Introduction to Data Science Capstone Project. From the initial write-up for this project, the goal was to determine if a proprietary ratings system for offense and defense in college football could determine the outcomes of future games when compared with the “spread” (the handicap that is given to a stronger team to level the playing field for even guessing). The idea was to have a winning percentage of more than 52.4% as this is considered the “industry average” for those who like to make these types of predictions.

The main source of data is comprised of the following types:

* The year in which a particular game was played.
* The week during the year in which the particular game was played.
* The team that was considered “away”, i.e. the team that had to travel for the game.
* The home team.
* The line or handicap for that particular game.
* Over/Under number – the predicted total points that both teams would score in that game (note that this will be used to do a different calculation.
* ATOR – Away Team Offensive Rating – Proprietary number based on away team’s past offensive performance.
* ATDR – Away Team Defensive Rating – Proprietary number based on away team’s past defensive performance.
* HTOR – Home Team Offensive Rating – Proprietary number based on home team’s past offensive performance.
* HTDR – Home Team Defensive Rating – Proprietary number based on home team’s past defensive performance.
* The final score of the game.
* Whether the game was played at a neutral site (information to possibly be used later).

Since the stated goal is to determine if the proprietary formulas could predict the outcome of the game, the first test was to prepare an initial calculation to study the formulas in of themselves and nothing else. A new data frame was developed to do this study only. The following elements were stripped away so this could be developed: year, week, over/under number and neutral site information. While most of the data in this set was tidy to begin with, there still needed to be a separation of the score as it was only listed as “Away Score-Home Score” format. This was separated at the dash into two distinct columns.

Now, new columns had to be added so that the formula calculation could be compared against the line in order to generate a predicted result. A simple method was used for comparison: (ATOR-HTOR) + (HTDR-ATDR). When the result returns a positive number, this will signify a preference for the away team and a negative number to favor the home team. This is analogous to the number representing the spread where a positive spread number would give the away team a handicap by that many points and a negative spread would give the home team a handicap by that value.

Once the number was calculated, a column was then added to show this value. Then, using this value, the prediction was made. If the formula predictor was more than the spread, the prediction was for the away team to be picked. If the formula predictor was less than the spread, the prediction was for the home team. A new column was added to the data frame to show which team was to be picked in the specific game.

Finally, results had to be tallied to determine success rate. A new column was added to show the actual winner of the game versus the spread. Then a final column was added to compare the cell in each prediction row versus the cell in the Bet.result row. The comparison would yield a “Win” or a “Loss” entry in the new column in to show the result of the prediction.

There are some limitations to the data based on what the industry uses to develop its own formulas for prediction. Some suggest that home-field advantage (a number of additional handicap points given to the home team based on past data suggesting that there is a 1-5 point advantage for the team playing at home) needs to be part of the data. Also, the formulas do not take into account prevalent injuries during games that could have an influence on the outcome of the rating in subsequent weeks. There is also no taking to account schedule strength. It is possible that Team A has better ratings for offensive and defense because it played teams that were not very good when compared to Team B which plays teams in the upper echelon of the league every week. This can worsen Team B’s ratings and therefore give an impression when comparing formulas, that Team A is statistically much better than Team B.

After doing the initial data analysis, it is shown that the prediction was correct in 2892 of 5594 games. This is a success rate of 51.6% over the span of seven years. There needs to be some improvement in the calculation to deem the use of this statistical set a success.

For the next steps, these calculations need to be analyzed. In each case, it is to be considered a milestone for the Capstone project:

* In the section on statistical analysis, it was required to do some statistics work on the data set to make some comparison of the data. For this project, a comparison was made on the proprietary offensive and defensive ratings versus the points scored/points allowed for each team. If the subsequent regression line equations (which had r-squared values ranging from .7-.8) were plugged into the rating to create a possible score and this was used to compare the two teams, there can be a determination of whether this proves predictive success.
* What were the results in specific years? Maybe there was one particular year that the model was bad and the other years were well above the 52.4% threshold. New data frames can be made and run the analysis to determine how consistent the model is.
* Can the time of the year make a difference? One would think that using the model for the first part of the football season would not be as successful because there is simply not enough data for the formulas to be successful. It would be a good hypothesis to say that the success rate should increase as the football season progresses through the year.
* Can there be any generalizations made about the success of predictions as it pertains to home-field advantage? Analysis can be done by grouping the predictor picks into five groups: Home favorites (in which the predictor model is more negative than the spread), home underdog (the predictor model is less positive than the spread), away favorite (the model is more positive than the spread), away underdog (the model is less negative than the spread) or neutral site games where both teams travel to the game site. This is mostly seen in bowl games and championship games at the end of the year, but there are also some neutral site games throughout the year.
* Could we look at standard deviation of the ratings to determine if calculations are skewed by one good result or one bad results and thus making it a less desirable predictor?

There are many other variables to look at regarding this type of analysis. For now, an analysis of these milestones can give a better indication of whether there are trends in the data. Remember, the goal is to get to a prediction rate of better than 52.4% on any subset of the data. If there is any advantage to selecting one set of data over another set, it could set a precedent in the industry.