Andrew Kerzak

2/10/2023

Econ 2824

Predicting the Winner of the 2023 Super Bowl

The Super Bowl has been one of the most popular sporting events in the United States for decades. The advent of legal sports gambling has only helped fuel the popularity surrounding the annual game. This year, it is predicted that over $16 billion dollars will be legally wagered on the game across the US. With that much money being placed on the game’s outcome, it is vitally important that the sports books setting the lines are accurately predicting the outcomes themselves as to not lose millions in payouts to its customers. For the average sports betting economist, it could be useful to predict their own model based on publicly available data prior to placing their own wagers.

To predict the winner of the 2023 Super Bowl, I believed that using historical data on past Super Bowl contestants regular season characteristics we can predict the winner on Sunday. I began searching for data by googling Super Bowl datasets. This search brough me to Kaggle where I could only find datasets filled with strings of information regarding the location of the game and other game facts not important to deciding the fate of the game. I would need to search for data that had information on the teams that could be related to the outcome of the game. Next, I revised my search to look for super bowl outcomes data, bringing me to the Pro Reference website where they keep a record of the super bowl outcomes. In this list, individual teams could be selected to view their statistics for the regular season and post season (prior to Super Bowl) in which they made the championship game. While this website provided all the information I was interested in, it did not provide an easy method to collect the data for analysis. This presented me with the biggest difficulty I came across finding data. After careful deliberation, I decided that I could either continue to search for the same data elsewhere in a more presentable form or bite the bullet collecting the data myself and creating a dataset from scratch. While time consuming, I elected to collect the data by hand to ensure that I had the features that I felt were important to the success of a team in the playoffs. Furthermore, to avoid additional practical issues with the data, I selected features that would not require extensive feature engineering, knowing the majority of my time would be spent on the data collection process. To ensure this, I selected league ranked variables that are already standardized in order to use the Multinomial Naive Bayes classifier. Additionally, I subset to the 2003-2022 Super Bowls and accompanying regular seasons. The time frame was selected due to the last NFL expansion taking place in 2002 and requiring additional standardization of the league ranked variables before and after the expansion if included. This allowed me to minimize the time I spent data cleaning and maximize the amount of time for the building of the dataset and the accompanying analysis.

When I had finished completing the dataset I had 20 observations, or one for each game over the 20 years of Super Bowls I subset. This included features on the league rank for total offense points scored, the league rank for total defense points allowed, the league rank for turnovers lost on offense, the league rank for the turnovers created on defense, and finally the Simple Rating System (SRS) for both the home and away team. The SRS ranks the league based on their strength of schedule and their points differential. In other words, the SRS judges a team based on the difficulty of their competition and how well they competed against competition (points differential). The SRS ranks teams based on their performance in the regular season. The offensive and defensive points are informative to a team’s ability to score and defend against scoring. While the turnovers show the teams ability to maintain possession and ultimately score.

In an ideal world, it would be useful to have a bigger dataset with more observations for the entire modern era of the NFL. Currently, the dataset is limited in scope without the need to standardize features to run the classification system. In its current form, additional features on the rank of the quarterback, special teams unit, and home/away splits could be useful if available for inclusion in the model. The inclusion of these additional features could fill in some of the grey areas of the game our model will inevitably miss. For example, if a team has a poor ranked kicker that has missed lots of field goals, they will have a poor special teams rank and have an impact on the outcome of the game.

By creating the dataset by hand, I had the opportunity to select and engineer the features that I wanted prior to loading in the data. I originally selected the Multinomial Naïve Bayes classification due to the ability to read discrete scaled features, the speed at which the classifier runs, and the ability to get an accurate prediction similar to other predictive models. Next, I considered the K-nearest-neighbor model. I chose to test this model because of its ability to classify based on the distance from similar observations. In the end, I selected the K-Nearest-Neighbor classifier because when compared to the other classifier the KNN achieves higher accuracy while producing a lower standard deviation. The K-Nearest-Neighbor classifier predicted the Kansas City Chiefs to win the super bowl with a model accuracy of 75% and a standard deviation of 0.13 at 1 neighbor. Additionally, we see that the KNN model standard error is minimized at 1 neighbor. Furthermore, the Naïve bayes classifier aligned with the KNN model and predicted the Kansas City Chiefs to win the Super Bowl. Because there are a low number of features, KNN does not suffer from the dimensionality problem and can easily find an example within the historical dataset.