Our goal is to create a model that will determine which college running backs are likely to be drafted into the National Football League (NFL). Our primary client would be NFL teams, who would be able to use this model to vastly narrow down the number of running backs they would need to more closely evaluate as they approach the NFL draft.

In order to create our model, we first needed to collect data on both NFL players and college players. Our NFL data came from the NFL web site, and since our model is meant to classify college players based on if they were able to make it into the NFL, we were mostly concerned with their names existing in the given data sets. Our college data came from the NCAA page that included the top 100 running backs from each year. While this does leave some players out, it is not unreasonable to assume that we may miss only a few corner cases as most of the players getting drafted will be among the best.

In order to collect the data from these web sites, we performed web scraping, where we wrote a program that automatically navigated through these web sites in order to collect our information and save it in a CSV file. Upon successfully collecting our data this way, we checked for missing data and converted any columns with numeric information into numbers, as the original import turned those columns into strings. We then performed an inner join on our NFL and college data so that we could determine which college players had made it into the NFL. This allowed us to create a column on the college Data Frame called ‘Made NFL’, which meant that we would only need to refer to the college Data Frame from this point onward.

There were other options for collecting our data. For the NFL data, we felt that using the NFL’s own web site would be one of the most reliable places to collect our data. While there are other databases that you can pay to access for more in-depth analysis of games for individual players, we felt that the NFL data was sufficient for our goals. For the college data, we may run into issues by only having statistics on the top 100 players. Most of the databases we discovered in our search were mostly interested in the records of specific schools as opposed to the information about the players. There were some databases that looked as though they may have had more information, but were hidden behind a paywall.

After performing some exploratory data analysis, we have found that average total yards for the season in college was the factor that was most correlated with making it into the NFL (labeled ‘Yds’ in the Data Frame). Having said that, we believe that our final model will include multiple factors in determining if a player is drafted into the NFL. There was only one statistic that was negatively correlated with making it into the NFL, and that was the average total negative yards for the season in college (labeled ‘Loss’ in the Data Frame).