Phil Robinson
SI 618 Final Project Report

Sports offers a uniquely interesting perspective into the beginnings of data science: aggregation of data makes intuitive sense, it is easy to describe findings to the average layperson, and findings offer a wide breadth of implementation. With the advent of highly specific data, data analytics is coming into a golden age, and this is seen just as well in sports as in other fields. College football stands in a highly describable state. Class standings, diversity of players, variety of conferences, and a deluge of trackable categories gives more than enough information to ask specific questions and come up with hopefully interesting answers.

There are four questions from the dataset that I found intriguing , and decided to further analyze. The first question: do freshmen running backs have a quicker improvement in their fumble rates, as compared to other classes? My initial guess would be that yes, freshmen would improve faster. Because freshmen have not played against the skillset of a college team, I figured they would be “sloppier” with the ball than their sophomore, junior, and senior colleagues. But as the season went on, and these freshmen had more experience playing against a more skilled defense, their fumble rate would improve more quickly than more experienced running backs, who already have at least one year experience playing against tougher defenses.

The second question: is there a linear relationship between the points a team is down by and the proportion of passes that a team makes, when narrowing the data to the fourth quarter. I would assume that as a team is down by more points, then they would throw the ball more. Because a pass play is a high-risk high-reward type of play, I would assume that as a team is down by more points, they would try more of these riskier plays to try and make a comeback. However, I also believed there would be an “inflection point”, where a lead becomes seemingly insurmountable, and the losing team just plays at an average pass/run play ratio.

The third question: does the Big Ten conference offer a uniquely defensive type of play, as compared to the other conferences? I will admit, this is the question that I had the least confidence in. As the past handful of season, the Big Ten has shown itself to be a very good defense conference, so more recent data would most likely show a significant finding. However, as I was not much of a sports fan in 2005, I was unsure of what the conferences defensive image was. This still seemed like a relevant question, in determining if the Big Ten had established itself as a defensively minded conference by 2005.

Finally, I wanted to test the old adage “Defense Wins Championships”, a ubiquitous term that never lingers far from sports reporters and commentators remarks. Can we prove one way or another if this is true? How important is defense as a marker for a team winning or not?

2. Data Source

The dataset that I have chosen to work with the final project is about the 2005 college football season, as supplied to cfbstats.com and coachbythenumbers.com.

http://coachesbythenumbers.com/wp-content/custom-php/sportsource-data-sample-advanced.zip

The dataset is a relational database of CSV files, all coming in to populate 17 tables. While some of the tables are straight forward (the game-statistics table gives the game code, attendance, and duration), others are much more complex (Player game statistics contains more than 60 columns) or contain a significant number of rows (the Play table, which details every play of every game contains more than 137,000 instances). The central table of the dataset tables is a play, with branching out relation sets diverging from there. In all, this is a very rich dataset, filled with lots of interpretable information.

One of my reservations about the dataset is that it is particularly old; where some datasets about other topics hold value across long periods of time, sports doesn't quite offer this, as player turnover is high. All of the teams since 2005 have had a complete roster turnover three or four times. Therefore, there is nothing intrinsically useful to be found in this specific dataset. However, what is useful is the practices and understandings that are to be gained. If I was to play fantasy football next year, and have access to a more recent dataset, all of these tools would be extremely useful in helping me make informed decisions about which players to select. There is more recent data, however, at a steep cost. Datasets of this detail more recent than 2005 have an introductory price of $350. So while I do wish I could have access to highly recent data, this dataset is useful as a proof-of-concept.

3. Methods

The two best aspects to this data was the variety of information that was tracked, but also the completeness of the data. Aside from one aspect, which is described in question 4, all of the data that I wanted access to was there.

Question 1: Framing of the data begins by merging the data into usable data frames. This is a recurring task through the project. While almost all of the data I wanted to look at was in the dataset, all of the questions that I ask involved bringing together separated data. In this first question, I had to bring the date of the game into the data frame that contains all of the plays in which I was looking at. I was then able to filter this data frame into data frames corresponding to class standing and then month in which the game was played (September - December). Once these data frames are created, I created two dictionaries, one a list of players by class standing, and another for a list of game codes in their respective months. I could then iterate through the dictionaries, calculating the fumble rate. This equation first finds the count of plays by class and month in which there was a fumble, and then divides it by the total run plays for the class and month. I multiplied this by 100 to get a percentage of plays in which a run resulted in a fumble per class per month. This data was then added to a new data frame, where the data could be visualized via seaborn.

One of the challenges that I faced when preparing this answer was to find an appropriate way to iterate through the data frames. A first attempt was to just hard code each instance of the fumble rate equation, specifically identifying which month and class the calculation was for. However, I decided to try and create a python dictionary of lists, which ended up working, so I was able to step through the data frames in an efficient manner.

Question 1 Analysis: When we plot the graph of the fumble rates per class per month, one of the first things that stands out is that my assumption was wrong in both aspects. For this graph, freshmen are red, sophomores are orange, juniors are yellow, and seniors are green. My first assumption was that freshmen in September would have the highest fumble rates among running backs, because of their inexperience at this higher level. However, we see that there really isn't any difference between fumble rates from class to class in September. While I do have data plotted for October, November is the real end goal I was looking for. In my initial prediction, I was expecting freshmen to have a higher fumble rate in September, and similar fumble rates for the classes by November. We see that by November, freshmen have the highest fumble rate; while not much different than sophomores and seniors, it is almost 50% higher than juniors. I am not taking October into consideration, as it is more of an intermediary month between the beginning and the end of the season.

This however, negates December. I am purposefully ignoring this month because I feel that it is not representative of college football as a whole. Instead of every team playing four times, like every other month (ignoring bye weeks), certain teams in 2005 would have one, maybe two games in December. These would be bowl games or potential makeups for games that had to be cancelled. The population size of December games is much smaller compare to other months. For the months of September, October, November, and December, there were 229, 259, 190, and 32 games, respectively.

Question 2: Filtering to the fourth quarter is the first step in finding the data that I wish to analyze. Then, we can further filter by looking at the plays in which the team on offense has fewer points than the team on defense. Preparing the data for the scatterplot ended up being much more difficult than I would have at first assumed. The data, while all there, is difficult to get to in an appropriate way. Unlike most datasets, where each of the rows has a unique identifying element, the data frame rows are all unique by the combination of play number, drive number and game number. Therefore, I had to go back to to the creation of a dictionary, where the key was the game code, and the value was a list of drive numbers. This allows me to iterate through each element of the drive list of each game, finding the pass rate for each drive and the points the team on offense was down by.

When I created the scatter plot and found the correlation, there was nothing of much use. Therefore, recreated the same data frame but added in the remaining time left, giving me the ability to filter through another dimension of the data.

Question 2 Analysis: While looking through the initial scatter plot, where I had not yet taken time left in the game into account, there was a very weak negative correlation, -0.109. This is not was I was expecting at all. While I didn't suspect a correlation of points down and pass ratio to be 1, I did expect it to be moderately positive. This is when I decided to alter my data frames, adding the time remaining dimension. By doing this, I can attempt to narrow down into more specific segments of the data, allowing more control of what characteristics I am looking at.

When segmenting through this new data, there was one iteration that showed some promise as to having more than a weak correlation. When a team is down by more than 14 points with 5-10 minutes left in the game, there is a correlation of -0.335. This is more along the lines of what I was expecting from the original hypothesis: at a certain point, a team deem themselves unable to make a comeback, and therefore, may take a more conservative approach: have a run play, where the clock is more likely to continue running, and where players are less likely to get hurt. From the original hypothesis, this is what I thought of as the inflection point, where the pass ratio would start to decrease, and with a moderate correlation, we can see this perhaps taking shape. However, as I continue to test different point differences and times remaining, there does not appear to be any other combination that diverges from 0 further than "greater than 14 points with 5-10 minutes left". There very well might exist a stronger correlation, but I believe I would have to greatly reduce the data to a very specific niche, potentially to the extent that the data is just very specific and no longer representative of true trends.

Question 3: This was the most straightforward of my questions, and was met with the least resistance. First, I needed to group team game statistics depending on which conference the team was in. To do this, I had to merge the data frame that dictates the conference a team is in with the team game statistics. I specified which columns were defensive features. Then I created two data frames, one that was specific for game statistics for a team in the Big Ten, and one that contained all of the other team game statistics, taking only the features that I had listed as defensive features. Then I iterate through the appropriate columns, asking about the p-value for the mean differences between the Big Ten and non Big Ten team game statistics.

Question 3 Analysis: In total there were eight exclusively defensive categories that I have wanted to test the Big Ten conference teams against to all the other teams. The first table are the p-values that were found for each of the eight categories. When comparing these values, the strongest p-value is for the category Tackle For Loss. Therefore, I wanted to visualize this via box-plot, to see what the data looked like. Based on this, the first thing to note would be that the p-value for this category, while the strongest, is not in itself inherently strong. With the value of 0.175, it is reasonable to assume that a visualizations wouldn't look too different at a quick glance, which is exactly what we see. Keep in mind, this is the number of tackles for loss, not the yards that were lost via a tackle for loss.

From this graph, we see the data does in fact show there is a slight difference in quantiles. While the Big Ten overall has a larger range of non-outlier data, the second quartile also shows a higher variation than compared to team outside of the Big Ten. But, because the p-value really isn't anything very significant, it would be a wasted effort to look further into the quantiles here, as we don't have significant evidence.

However, for one final analysis, I integrated more features into the analysis, to further compare. This included things that I would consider less defense specific, but still within the defense category. When I performed this analysis, I found a category that I had initially left out of the analysis that I had somehow overlooked: Pass Interceptions. This feature has a p-value of 0.052, very close to a classic significant p-value. When I created the box plot for this category, expecting the Big Ten to be a very good interception team, I actually found quite the opposite. The Big Ten in 2005 was actually a very poor defensive interception conference, and the box plot suggests this to be true. We can see that the median of Big Ten teams is an entire interception below teams for all other conferences, and the fourth quartiles are not even remotely similar.

Question 4: Answering this question was by far the most challenging. To begin, I immediately faced the challenge of the team game statistics table having 65 features, but wins or loss not being one of them. In order to properly generate this label, I needed to find the table that had the final score of each game, which unfortunately doesn't exist in this dataset. However, I did have each play of every game, which lists the team on offense, the team on defense, and the points for both of these, as we saw in question 2. In order to get the "final score" of each game, I generated a new data frame, containing the last play of every game, and therefore, the final score of each game. Merging this final play to each to the corresponding team game statistic data frame, I was then able to write a function that determines if the team won or lost the game.

Next, I filtered the merged data frame back to only the 65 game statistic features and the win-or-loss label. Utilizing the tools we learned for a classification task, I used a Random Forest Classifier to determine what a win or a loss would be in the training data.

While I won't spend time talking about it here, I was able to determine that a 0.6/0.4 split in train/test data was most likely the appropriate proportion for the Classifier. This can be seen in my submitted code.

Once the classification was completed, I was able to attempt to answer my underlying question: does defense win championships. To do this, I created 4 groups of features: offense, defense, kicking, and penalties. I was able to sum the weights of these features, and generate a pie chart, representing the weights of these feature categories in determining a win or a loss.

Question 4 Analysis: After I had aggregated the weights of the features, Kicks was the set of features which ended up being the most important set for determining a win or loss label, coming in at 59.8% of label weight being attributed to kicking statistics. This was, to say the least, unexpected. I was unsure about how true this could really be, and upon further investigation, the feature of Kickoff Yard commanded more than 30% of the label weight. The first graph here is for the four groups, representing the weight of each group to the total label. Because this had created such a lopsided representation, I decided to retrain the model, with the same procedure, but instead I removed the Kickoff Yard feature from the model. When I did this, I found data, that while still showing kicks to be an important part of the label, showed much more intuitive data. The only caveat is the removal of data that seems to carry quite a bit of importance. However, when the accuracy of this new model was tested, it was shown to have the same accuracy as when Kickoff Yard was in the model. Therefore, I felt it was within reason to remove the feature for exploratory purposes. This raises an important question, however: if this feature carries such a significant weight, why does the removal of this feature not degrade the accuracy? There are two possible answers that I currently see. The first is that I had not used enough trees to to properly handle the features in an appropriate way. In order to test this, I created an iterative loop, changing the number of trees and measuring the corresponding accuracy. This had almost no impact, so for now, I go with my other assumption, that the weight was able to distribute to the other features without a decrease in the total classification accuracy.

When I removed the Kickoff Yard from the feature set, the data becomes more what would be expected, although still not my first assumption. By removing this feature, offense becomes the most important feature set, with 41.6% of the label weight being attributed to the label. Following this is kicks, defense, and penalties.

This really was to get to the question: does defense win championships. Because the feature set is really looking to predict with group is the most important for the classification, it appears that for 2005 college football, offense was the most important aspect for determining a win.

I suppose the phrase may need a bit of tweaking.

Conclusion:

In all, this was an enjoyable dataset to work with. What I enjoyed most about this dataset was just giving myself the opportunity to explore my own creativity with the data. Seeing the features that I had available, and thinking about new and unique ways to try and digest the data was one of the most enjoyable parts of this semester for me! I found myself really connecting almost all of the concepts from the semester into this project, and implementing concepts into actual findings. While most end of the semester projects are a pain to work on, this really was a very enjoyable experience! I had my fair share of battles with the data, but I have cemented the concepts here, and cannot wait to see how I apply these in my future.