Data Mining in Sports

Chris Heinrich

Advisor: Richard O'Keefe

Department of Computer Science
University of Otago
Dunedin, New Zealand

cheinrich@cs.otago.ac.nz

July 19, 2013

Contents

List of Figures

List of Tables

Abstract

This project explores data mining in the ever changing sports world and the exciting challenges this provides. American Football is the sport we decided to focus on because of how often the game changes and the available data. This project attempts to study and replicate previous sports prediction models and measure how the changes in the sport affects model performance over time.

The report is broken into six chapters, which will help you understand the process we have taken. It will give you an overview of American Football and some of the available, relevant journal articles. Then a brief summary of the available data will be given, followed by our main theme of "change". The report will end with our success at model replication and goals for semester two.

1 Introduction

1.1 Project Outline

This project had 3 main first semester goals:

1. Explore data mining in sports
2. Replicate a previous model
3. Study sport over time and try to find changes that have occurred. Explore what this means to past models and how to spot this change.

I began my exploration of data mining in sports with an extensive literature review. I was mainly trying to find what models people have created and what data they have trained/tested it on. This part also featured learning the R programming language, as well as basic data mining techniques.

Replicating a previous model meant that we had to acquire our own data about the sport. This consisted of looking at available datasets and selecting the best option. Doing the actual replication allowed me to gain experience with the R programming language. It also ended up serving as a check on our dataset because it was compared against another dataset.

The last goal of finding and spotting change that has occured in the sport has turned out to be the most interesting part of the project. This topic hasn't been explored by others, which makes it more exciting. This deals with the different types of changes that can affect the way the sport is played. This change has an effect on the game and we want to explore it's effect on past model performance and if there is any way to account for it.

1.2 Brief Overview of American Football

American Football is a game played between two teams with the goal of trying to score more points than the other team. The game is broken down into 4 (15 minute) quarters with a 30-minute break (half time) after the 2nd quarter. At the start of the game, a coin toss determines who kicks the ball to the other team to begin the game. At any point in a game, one team is trying to score (offense) and the other team is trying to prevent the other team from scoring (defense). Once the receiving team (offense) has possession, they have to gain 10 yards in 4 plays or else the other team gets the ball. The team with possession of the ball's goal is to get the ball to their opponent's goal line (called a touchdown). They are awarded 6 points for this and are also awarded 1 attempt to either kick an extra point (1 point) or try to get across their goal line again when starting 2 yards out (2 points).

To score a touchdown, an offense has two main ways of gaining yards. Gaining yards means that the offense has run a play and moved the ball closer to the other team's end zone. To gain yards teams can either run the ball (rushing) or throw the ball (passing).

These two strategies have their pros/cons and are often used in some combination to try to score a touchdown.

To prevent a touchdown, the defense will try to tackle the offensive player before they are able to gain 10 yards from where their offense started. If they are able to prevent the offense from gaining 10 yards in 3 plays, the offense will usually kick the ball to the other team to give them the longest possible yardage to score a touchdown.

Kicking is an underrated and important aspect of football. It can often be the difference between good teams and can change the momentum of a game. When an offense has failed to gain 10 yards in 3 plays, they will usually opt to kick the ball to the other team (Punting). However, if they are close (usually around 30 yards or less from a touchdown) they can opt to kick a field goal (3 points). If the kick is successful, the team that kicked will be awarded 3 points. However, if they miss the field goal, the other team will get the ball at wherever spot on the field the field goal was kicked at.

The league that makes up all the teams in professional american football is called the National Football League (NFL). It is comprised of 32 teams split into 2 conferences (AFC & NFC). These conferences are then divided up into 4 divisions (East, West, North, South), each containing 4 teams. Teams will play 6 of their 16 games against other teams in their division. Teams will typically play only 4 games against teams from the other conference.

The NFL has a committee called the “Competition Committtee” [13], which consists of eight members. These members are either head coaches or involved in upper level management of NFL teams. If a member of the committee loses their job with the NFL team, they lose their membership to the committee as well. This committee meets once a year to oversee the competition of the league and suggest rule changes. These rule changes are then voted on by the team owners.

2 Literature Survey

2.1 “Hidden Game of Football”

The book, “The Hidden Game of Football” [7], provides many interesting perspectives about American Football. It discusses different changes that have occurred to the game and these changes helped shape the direction of my project. There are at least three kinds of change: Changes in reporting (what is noticed and recorded), changes in the rules (what is allowed), and changes in play (what is done by teams on the field). These changes will have different effects on past models and therefore must be viewed separately (Expanded upon in Chapter 4).

This book also brought to our attention the fact that the NFL teams are looking at the same data we are. This means that the NFL as a system is self-aware or it contains people who are looking at the same data we are. This fact will help explain trends we see in graphs and model performance. These two points of the system being “self-aware” and change throughout the game will be expanded and explained more thoroughly later in the report.

2.2 Journal Review

Publication Date	Year(s) Data is from	What Techniques	Reference Number
1991	1981, 1983-1984	Logistic Regression	[20]
1997	1995-1997	Logistic Regression	[9]
1998	1988-1993	State-Space Model / Bayesian	[11]
2003	2003	Back Propagation & Multi Perceptron	[12]
2006	1992-2001	Logistic Regression	[10]
2007	2004-2006	Linear, Multiple, and Logistical Regression	[18]
2008	2004-2007	Multiple Regression Models	[22]
2009	1970-2006	Logistic Regression	[6]
2010	2000-2009	Gaussian Process Model	[21]
2010	2009-2010	Logistic Regression, Support Vector and AdaBoost	[14]
2011	2003-2010	Artificial Neural Network	[5]

Table 1: Overview of models used by articles found during literature review

Table 1 shows a brief overview of the important information gained during the literature review process. A majority of the papers used logistic regression in their model and this was for good reason. Logistic regression is often advantagous when the variable being predicted is binary [17]. Therefore, these authors saw that in American Football a team can either win or lose. They were ignoring the possibility of a tie because they are infrequent. A tie was also likely not taken into account because a lot of these papers focused on predicting American Football games to gain an advantage on the sportbooks (gambling agencies). With sportbooks, ties are essentially counted as a loss for the gambler.

The articles referenced in Table 1 were all trying to accomplish the task of predicting the outcome of an American Football game. What was unique to each of the journals was the date of the data they decided to use to build/test the model. We became particularly interested in this data when we started looking for change in the data over time (Chapter 4).

The journal review process also helped us gain perspective on how well a "good" model performs. In 1998, using their unique state-space model, [11] achieved a 58% prediction success rate for second part of the 1993 season. In 2003, [12] claimed a 75% success rate, however, this used only two weeks of games as test data. [10] claimed a success rate of 51% and 57% for 2001 and 2002, respectively. In 2009, [6], Blundell achieved a 62% success rate with a logistic model that he improved to 65%. [21], using 2008-2009 testing data, achieved an impressive success rate of 64%. Some journals have been left out because they factored in gambling to the success of their model, which we don't plan on doing.

The final part of the literature review process was to explore data mining techniques. This included learning about the programming language R as I had not used it before. The book, "R and Data Mining: Examples and Case Studies" [24], provided tutorials and case studies to help me gain a basic understanding of the language. Two books ([23] & [8]) provided additional knowledge on how to use the R libraries and how to understand graph trends and what they could mean. The information provided on data visualization ([8]) would prove vital in trying to find the change that occurs over time.

2.3 Unasked Question

"The Hidden Game of Football" [7] exposed us to this topic of change in American Football. However, throughout the literature review process, I found that no one has published anything relating to this change over time. Many journals will talk about the model they chose and how well it runs, but they don't deal with change throughout the years and how that could affect their model. So while they may have a model that performs well during a certain time period, it could perform badly during another time period. This violates a common data mining assumption that you can train/test the model on data from any time period and it will perform similarly.

This change in the game can lead us to three outcomes regarding previous models:

1. We can improve upon a previous model because of new data
2. The previous model will perform worse on new data
3. The performance will stay the same because the new data is consistent with the old data

We can improve upon a previous model because of new algorithms and statistics that weren't available when the model was created. The previous model will perform worse on new data because change has occurred in the game. For example, this could mean that more yards are gained because of new penalties designed to protect the offensive players. This increase in yardage would not be in previous models, therefore, potentially affecting their performance. Lastly, the performance of the model could stay the same. This would be because, using the last example, the model doesn't use yardage in determining the outcome of the game. Therefore, the change in yards gained doesn't affect the model or its performance.

It's unknown why this aspect of American football has been overlooked, but it creates an interesting opportunity to research something that hasn't been published.

3 What We Bought

3.1 Two Available Datasets

The first task to accomplish our goal was to find out what data is available. The first data source I considered was Pro-football-reference.com [16]. Numerous papers in the literature review used this site, however, two main problems arose. Firstly, the site wasn't coded well and it would be difficult to run an HTML parser to collect all the information necessary. Secondly, the site's terms of use statement prohibited pinging the site repeatedly. These concerns led me to look for other data sources.

A second option was found, nfldata.com [15]. However, this site required you to buy their data. This option had the benefit of having fewer missing values, more attributes (228 vs. 41), better structure (including a data dictionary explaining all attributes) and favorable terms of use. Expanding on these benefits, the free (Pro-football-reference.com) option didn't include playoff data. This is an extra 10 games that were included in only the paid dataset. Because of these benefits, as well as the time saved from having to write an HTML parser, we decided to buy the this dataset, which I shall call the "paid" dataset.

3.2 Data Summary

This "paid" dataset [15] came in 6 files, each containing information regarding different aspects of American football from 1985-2012.

Table Name	Number of Rows	Number of Columns	Number of Unique Columns
TeamGame	14,088	228	107
Schedule	7,044	8	8
Game	7,044	203	107
Standing	845	17	17
TeamSeason	845	218	105
Team	35	6	6

Table 2: Comparision between rows, columns and unique columns between different files in the paid dataset

Table 2 shows the number of unique columns which is a way of showing which variables are only counted once instead of twice. They are counted twice because there are two records (rows) for every game. This is done because every team has one record when the record is in their "perspective" and the other when they are the "Opponent" (in another team's perspective). So the unique columns shows how much information is gained about that team for that record.

TeamGame

This file consisted of each game twice (teams are placed in different "perspectives").

This file contains the maximum amount of information available from this dataset for each game. This is the file we used during model replication.

Schedule
This file contains the matchups between teams for every week between 1985-2012. It also includes gambling information (e.g. Point Spread and Over/Under) about the matchup.

Game
This file is similar to TeamGame, but it only has each game once (instead of twice). It contains all the same unique columns as TeamGame with half the number of rows.

Standing
This file contains the end of the year standings for every division in the NFL. You can tell who won every division and various other statistics about how well/poorly each team's season went.

TeamSeason
This file shows various information about each team at the end of the season. It has various statistics (e.g. Points scored and points scored against) to show how well the team did on the field for that season.

Team
This file shows for every team, what division and conference that team is in. It also shows the full team name and what city the team is from.

3.3 Missing Data/Data Not Acquired

There is some missing data in the dataset. Roughly 1/4 of it is missing because new statistics began being measured in 2001. Before 2001, the cells under those columns for previous years are left blank. This restricts the number of variables we can put into older models. This is one of the changes mentioned earlier. However, just because new ways of measuring the game were introduced, doesn't mean that previous models performance will suffer. It's just another way to observe the game and create more complex models.

We haven't acquired any player datasets, so we can't measure individual player importance to a team or model. This type of information would allow us to build more complex models, as well as look at the change that occurs when players switch teams.

4 A Changing Game

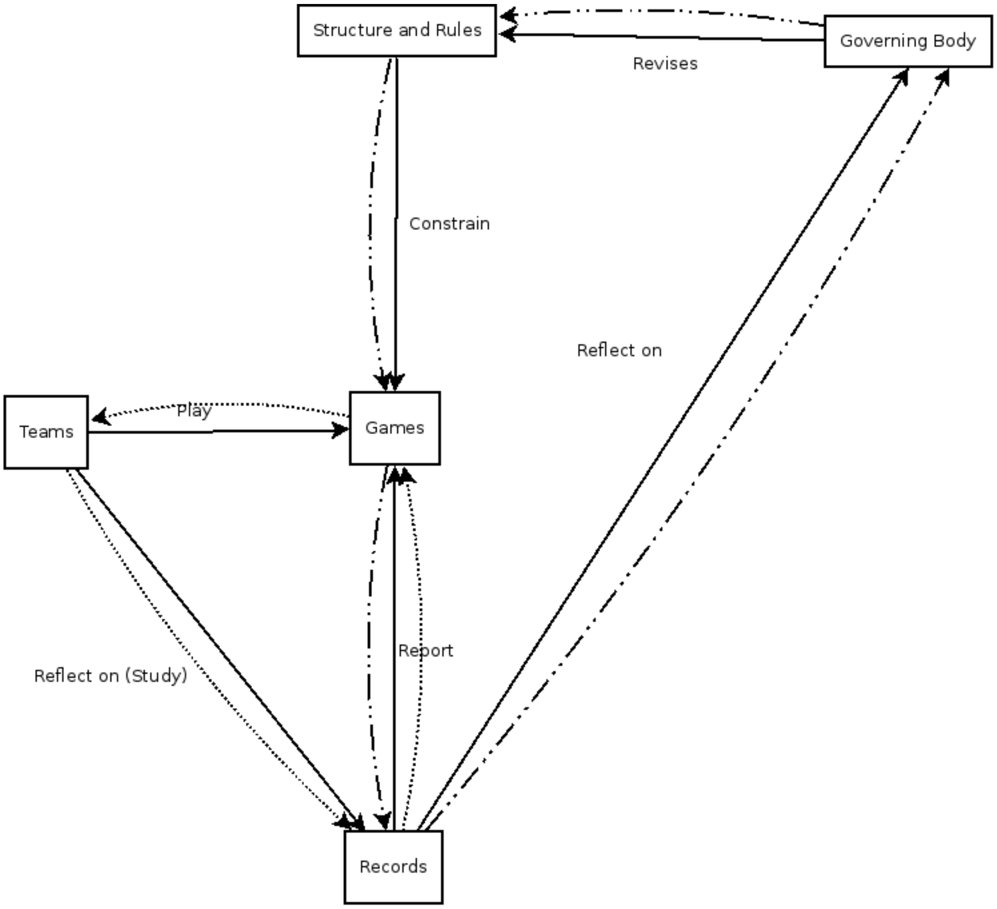

Figure 1: Diagram showing process of change throughout a sport

4.1 Understanding Change

Figure 1 shows the process of change and how it occurs in sports. There are two feedback loops. One is done by teams between weeks of the season (dotted line loop) and the other is done by the governing body at the end of each season (semi-dotted line loop). This distinction is important when looking for trends or change and noticing where it occurs.

After the teams have played a game, they go back and analyze what happened. One of the things they will look at is the records or statistics of the game. They are looking for statistics that they performed poorly on and thinking about changing their strategies to improve on them. This type of change will reflect week to week and should occur league-wide. The changes that the coaches make to their team are normally small changes. They can't risk changing all their strategies or their team will have a hard time adjusting to them. This means that the changes they implement should have a smoother slope.

After the season, a governing body (Competition Committee) convenes and they look at many different aspects of the sport. Issues ranging from player safety to rule changes are discussed at this meeting. Since these changes to the game are made at the end of the season, the effect that these will have on the sport can not be determined until the next season. This will result in a step change instead of the smoother slope above. This is because typically these changes have a big effect on how the game is played and it takes time for teams to react. A test has been developed to "detect and estimate gradual changes of the mean value" [4] over time. This test was developed to help with climate change data, but will potentially help with our data. It could help because this test could be included in models to spot change and act accordingly.

A key difference between these two types of change is how predictable each one is. If you examine a team's statistics, you will be able to see what they are poor at and predict that they will learn and improve upon it. However, with changes that are implemented by the governing body, you aren't able to predict what issues they think are most important to correct at the given time. The game could be facing many issues and the governing body would have to decide which ones demand changes and which ones can wait.

The most important thing about Figure 1 and the change process is that it can be applied to most large team sports, not just American Football. Most large team sports follow a similar structure and we believe we will be able to spot these changes in other sports. There will be subtle differences, but the main themes should remain the same. There will always be teams examining themselves and trying to improve on their deficiencies (which will result in a smooth slope) and changes that occur year-to-year at a league level. This is why we believe if we learn how to find change in American Football, we can use similiar techniques to find it in any large team sport.

4.2 Rule Changes

American Football is constantly evolving and rule changes are one way this evolution occurs. Rule changes occur after every season. They are done to address current issues within many aspects of the game. They can range from allowing teams more time between plays to introducing new ways to score more points. These rules changes don't have to be balanced and can affect teams differently. For instance, if a rule introduces penalties to protect the offensive players, then those players have received extra protections and therefore, an advantage over defensive players.

This means that past models can lose their accuracy depending on the variables the model depends on. However, if the change affects all team equally, then it will result in a shift for the whole model. This means that the model still has the same performance, but needs to be tweaked to account for this shift.

Some examples of rule changes that have occurred include [1]:

- 1977 - Defenders can only make contact with offensive pass catchers (wide receievers) once. This was done to “open up” the passing game.
- 1988 - Time inbetween plays was increased from 30 seconds to 45 seconds.
- 1994 - The 2 point conversion is introduced. This gives teams the option of opting to try to score 2 points after a touchdown instead of just kicking a field goal for 1 point. Defensive players are now given a penalty if they move across the line of scrimmage and cause an offensive player to move. Kickoffs were also moved back 5 yards.
- 1995 - Quarterbacks (Offensive leader) now have a radio in their helmet to communicate with their coaches on the sidelines.
- 1999 - Instant Replay is brought in. This allows each team to challenge two referee rulings per game.

4.3 Finding Change

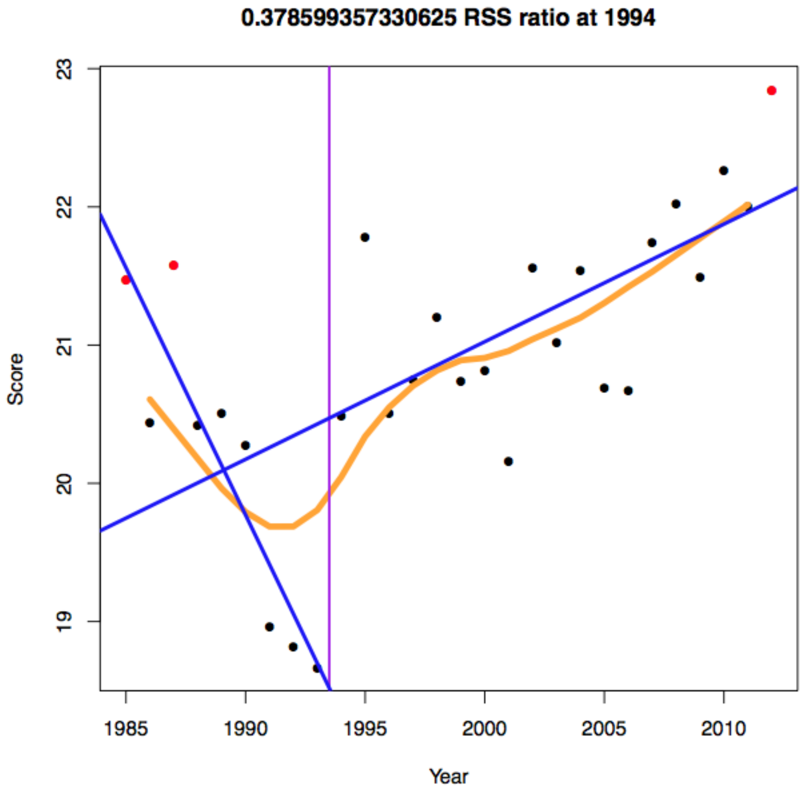

Figure 2: Example where change has been found

Figure 2 is an example of what we believe to be change occurring in the game. The orange line is the mean of the variable "Score" for each year. The two dark, thicker blue lines are lines of best-fit (one in each direction) through the data. The thin blue line is where the program estimates the change took place. The mean of "Score" was on a steady decline until around 1993-1994 when it began increasing again. We believe something happened around this time to cause this change that we see to occur. This is what we have to explore and try to find out what caused this change.

To find instances of where change has occurred, a program was written called "find.breaks.r" (used above). This R program's goal is to find instances in the paid dataset where change has occurred for a particular variable over time. The code consisted of plotting the mean of each variable for each year and then determining when change has occurred. To determine when change has occurred, we programmed it to find a break point in each attribute that gives the best least squares fit. This is done by looking at a variable for each year and splitting the data into two groups: Data before that year and data after that year. A linear model for each side of data is created (thick blue line). Then we compute each side's residual sum of squares which tells how good of a fit the model is to the data. We then add these two values together to measure how well of a fit that years model is to the data. This is done for every year in the dataset and we perform a linear search to find the year with least sum of squares (best fit). It then looks at if there was a change in height, slope, or both. Basically, we are looking for an improvement over fitting just one line. This was done for all attributes in the dataset.

The result of this program was that we now had some specific examples of where/when change occurred in the paid dataset. There were a number of examples that had the change occurring around 1994 (Figure 2 on page 11 & Figure 4 on page 17). We have begun looking at why this change occured. Was it a rule change, like moving back kickoffs (1994)? Moving back the kickoffs would give the offensive team better starting field position and less distance to score a touchdown.

Another attribute that showed interesting change was "OffensiveYards" (Figure 3 on page 17). "OffensiveYards" was in a decline before 1990 and then it has been increasing ever since. We are exploring whether this was an effect of the rule change that occured in 1988 which increased the time between plays. Did the teams figure out that they can use that time to see what formation the defensive players are in and how to take advantage? Another question that will need to explored further is, Did these changes affect all teams equally or were some teams affected more than others?

4.4 Self-Aware System

This idea of the NFL being self-aware was an important one to understand because it could explain the trends seen in graphs and model performance over time. Basically, teams will see the trends that are occurring and the attributes most important in their models to predict games. Because of this, they are able to learn and adjust the way they play the game to increase their odds of winning the game according to a model.

We hypothesize that these slopes and trends we see in the attribute means are the teams learning. The question this leads to is, What does learning in the system do to the data mining? Does this learning slope render previous models using the affected attributes irrelevant? This is where we are now and we have to figure out what learning does to previous models and how to spot this learning.

5 Model Replication

5.1 Introduction

The last objective for semester one was to replicate a previous model. During the literature review process, I was able to find numerous suitable models that could be replicated. These journal articles either provided their code or clearly showed their model and how it was formed.

We decided to use a very simple model ([2] & [3]) for our first replication. The model predicted the result of a game between two teams by determining what each team's score would be. It did this by fitting a linear function to the attributes "Score" and "OffensiveYards". Then it would calculate the two teams average "OffensiveYards" up to the that point in the season (like a running average) and determine how many points they are estimated to score (according to the linear function calculated earlier). Lastly, the team with the higher score is subtracted from the team with the lower score. This is done so you can tell how much of a favorite one team is from another.

Replicating a model served three purposes. Firstly, it provided us a check on our paid data. This is because the author of the model used free data [19] and provided the link to download that data. We would replicate the model using our paid data and see if the results matched up. We also would run summaries on corresponding attributes in the two datasets and see if there are any inconsistencies. Secondly, it allowed me to get my feet wet with the R programming language. It allowed me to gain some confidence and experience with R in a relatively controlled environment. Lastly, it showed me how models are created. The author provided a clear walkthrough of the steps he took in creating the model. This will help me during second semester, when I will create my own model.

5.2 Preprocessing

Before we began replicating the model, we first examined the datasets. I wrote a piece of a code that compared the summary (function in R) between every corresponding pair of variables in the two datasets. Corresponding pair of variables means that, for example, "Score" in the paid dataset was compared against "Score" in the free dataset. The summary function provides familiar statistics such as min, max, mean, median, 1st quartile, and 3rd quartile. Using this information, I went through every corresponding pair of summaries and looked for any differences.

There were differences. When I investigated, I found that they were caused by different patterns of missing data, not by recorded data disagreeing. The free data has many more missing items, with no explanation. The missing values didn't end up affecting the mean very much, but it was still not as complete as it could have been.

5.3 Results/Conclusion

Model	Prediction Success	Prediction Success (games with 3+ point favorites only)
Original Model	48.4%	50.5%
Our Replication Model	48.24%	50.8%

Table 3: Comparision between results of our model and the model we are trying to replicate

We were able to replicate the model's results. This was reassuring because it meant that our paid data was consistent with the free data. The slight differences are what we believe to be the missing data in the free dataset. There will be no similar way to test all the attributes throughout our paid dataset, but at least we were able to check the statistics that are "mainstream" or readily used.

6 Second Semester

6.1 Goals

1. Replicate complex model
   I want to replicate a more complex model to give me ideas about what type of model I can create. Also this will be a model that we can use recent data on and see what kind of changes has occurred.

2. Use new data on past models
   Use new (recent) data on past models and look for any changes that have occurred in its performance.

3. Create own model
   I would like to create my own model using the latest algorithms/technology. I will have to develop a method for determining which attributes to include. I hope to incorporate some way of finding when change has occurred and (ambitiously) have the model correct for this.

4. Compare results to other sports
   I would like to try to find change in other sports and see if the results we find in American Football are applicable to other sports.

6.2 Process

	Old Data	New Data
Old Model		
New Model		
Better Model (Fitted to Old Data)		
Better Model (Fitted to New Data)		

Table 4: Table to complete for second semester

Table 4 shows the table I hope to complete for second semester. This table will look at both different models and different data. Their performances will be compared to spot possible changes.

- The "Old Model" will be a model that we have replicated (using old data). This will then be fed new data and the performance between the two will be able to be compared.

- The "New Model" will be the one that we have created. This model will try to predict NFL games across various years. A way to determine when change has occurred will need to be incorporated. This new model will be measured both with old data and new data.

- The "Better Model (Fitted to Old Data)" means we will try to improve upon the old model to try to improve it's performance (for old data). This means that we could add/remove variables or implement new/recent algorithms to improve upon the performance for old data. This will then be tested against the new data.

- The “Better Model (Fitted to New Data)” means that we will try to improve upon the new (created) model when it is fitted specifically for new data. Again, we will tinker with the structure to try to improve the performance for new data only. This will be compared against its performance with old data.

6.3 Conclusion

I am pleased with the amount of progress that has been made and am looking forward to the challenges of second semester. I am glad to have replicated a previous model as it makes the task of creating my own model seem doable. Doing something that hasn’t been done before, examining change in American Football, has been exciting and motivating. I am hopeful that we will be able to find useful information in this regard and contribute to the scientific community.

A Figures

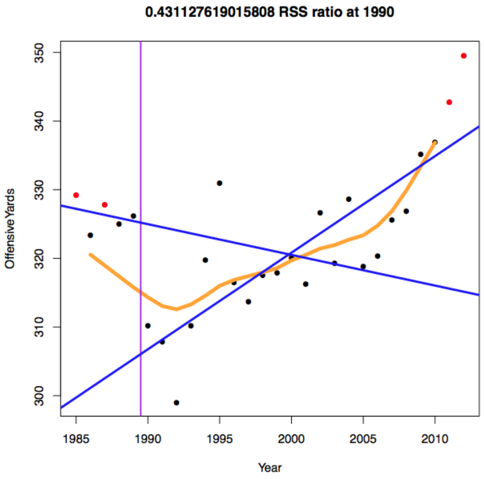

Figure 3: Graph showing the attribute “Offensive Yards” over time

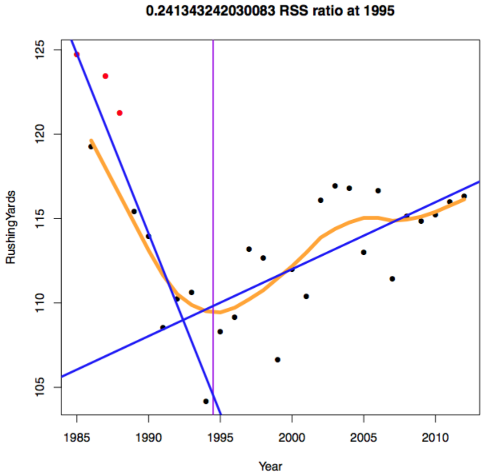

Figure 4: Graph showing the attribute “Rushing Yards” over time

References

[1] History of NFL Rules. http://www.steelersfever.com/nfl_history_of_rules.html.

[2] Model Building Part 1. http://docs.google.com/document/pub?id=1ePpq9vS7nBOXzN8RZaDhkgkOAe0yuA8F1tidrwQpPbw, 2013.

[3] Model Building Part 2. http://docs.google.com/document/pub?id=1R8jxFbtByNkmKjVkYVnIyCQjaFoCfZ9hWcU9ELxjFTc, 2013.

[4] Hans Alexandersson and Anders Moberg. Homogenization of Swedish Temperature Data. Part 1: Homogeneity test for Linear Trends. *International Journal of Climatology*, 17(1):25–34, 1997.

[5] Andrew D Blaikie, Gabriel J Abud, John A David, and R Drew Pasteur. NFL & NCAA Football Prediction using Artificial Neural Networks. 2011.

[6] Jack Blundell. *Numerical Algorithms for Predicting Sports Results*. PhD thesis, University of Leeds, School of Computer Studies, 2009.

[7] Bob Carroll, Pete Palmer, John Thorn, and David Pietrusza. *The Hidden Game of Football: The next edition*. Total Sports, 1998.

[8] David Chambers. *Software for Data Analysis*. Springer, 2008.

[9] David N. DeJong. Using Past Performance to Predict NFL Outcomes: A Chartist Approach, 1997.

[10] Kevin Gimpel. Beating the NFL Football Point Spread.

[11] Mark E. Glickman and Hal S. Stern. A State-Space Model for National Football League Scores. *Journal of the American Statistical Association*, 93:25–35, 1998.

[12] Joshua Kahn. Neural Network Prediction of NFL Football Games, 2003.

[13] National Football League. Jeff Fisher, Mark Murphy & Ken Whisenhunt Named to Competition Committee. http://www.nflevolution.com/wordpress/wp-content/uploads/2012/08/new-nfl-competition-committee-assignments-2-14-12.pdf, 2012.

[14] Brian Liu and Patrick Lai. Beating the NCAA Football Point Spread. 2010.

[15] NFLdata LLC. Team Historical Database. http://www.nfldata.com, 2013.

[16] Sports Reference LLC. Pro Football Statistics and History. http://www.pro-football-reference.com/, 2013.

[17] H Dunham Margaret. Data Mining Introductory and Advanced Topics. 2003.

[18] Everson Phil. Teaching Regression using American Football Scores.

[19] Warren Repole. Sunshine Forecast Downloadable Data Files. `http://www.repole.com/sun4cast/data.html`, 2013.

[20] Hal Stern. On the Probability of Winning a Football Game. *The American Statistician*, 45(3):179–183, 1991.

[21] Jim Warner. Predicting Margin of Victory in NFL Games: Machine Learning vs. the Las Vegas Line. 2010.

[22] Brady T West and Madhur Lamsal. A New Application of Linear Modeling in the Prediction of College Football Bowl Outcomes and the Development of Team Ratings. *Journal of Quantitative Analysis in Sports*, 4(3):1–19, 2008.

[23] Nathan Yau. *Visualize This*. Wiley, 2011.

[24] Yanchang Zhao. *R and Data Mining: Examples and Case studies*. Academic, 2013.