The running back is ready to run the ball to get the first down, but is there enough information from the National Football League to predict that the play will be successful?

Christopher Rowe

Introduction

In 1982, the public was introduced to Bill James and a group of people who have created "Sabermetrics".[1] Sabermetrics introduced advanced baseball statistics that help find better players. These statistics include the batter's On-Base Percentage (OBP), the pitcher's Walks plus Hits per Innings Pitched (WHIP), and even the individual player's Wins Above Replacement (WAR). In 2001, Major League Baseball (MLB) fans saw the effectiveness of statistics in baseball with the Oakland Athletics' run into the playoffs, including their historic 20 game win streak. This story was popularized with the 2011 movie *Moneyball*, based on the 2003 book.

It's not just baseball that has seen an increase. In 2015, the National Football League (NFL) started collecting data for sports analysis. The recent deal with Amazon,[2] not only to allow Thursday Night Football games to be broadcast on Amazon's streaming service Twitch, but also allowing the use of Amazon Web Service (AWS) to utilize and optimize the data for statistical analysis in their "Next Gen Stats" program. It is estimated that by 2022, the market for big data in sports could be worth around four billion dollars.[3] And the data isn't limited to just on the winning on the field, but also fan interactions and purchases as well.

What this project will do is apply the data, all running plays throughout three years of the regular season provided by the NFL, to investigate whether

1. Phil Birnbaum, "A Guide to Sabermetric Research," `https://sabr.org/sabermetrics/single-page`.

2. Jeffrey Dastin and Anirban Paul, "Amazon, NFL reach $130 million streaming deal for Thursday night games: source," 2018, `https://www.reuters.com/article/us-nfl-amazon-com/amazon-nfl-reach-130-million-streaming-deal-for-thursday-night-games-source-idUSKBN1HX3EP`.

3. Abhas Ricky, "How Data Analysis In Sports Is Changing The Game," 2019, `https://www.forbes.com/sites/forbestechcouncil/2019/01/31/how-data-analysis-in-sports-is-changing-the-game/`.

or not we have sufficient information predict the running back's ability to successfully make a first down. Based on this, the project will use several methods of regression analysis to create prediction models, to see if we can calculate the probability, and which factors contribute the most, for successfully making the first down from a running play.

The objective of a football game is to receive the most points in the allotted time. Points you can earn in one possession include scoring a touchdown. Each team consists of 11 players on each side. One team is delegated to have possession of the ball, the other is on defence. The team that has the ball has the task to get the football to the end zone for the touchdown. A standard length of a football field is 100 yards. Each play (called a "down") starts with the ball, which is placed at the yard line where the last play was finished and is hiked by the center to the quarterback. The quarterback will then either throw the ball to a teammate, called a *passing play*, or hand the ball to a running back, called a *running play*. The objective of the player who has the ball to gain as many yards before a defensive player tackles him to the ground. At the start of the possession, (or "drive"), a marker is placed 10 yards from the spot of the ball, which acts as a checkpoint. The possessing team has up to 4 attempts to make it past that 10 yard marker. If they succeed, that's called a "first down", and the drive continues with a fresh set of downs and a new target of 10 yards to obtain. On the final attempt, (referred to a "fourth down"), they can choose to punt the ball so it goes further for the opposing team, (who then takes possession), or they can choose to go for the first down. Failure to obtain the first down will result in the defending team to take over as the possessing team on where the ball was finally spotted.

Literature Review

Sports predictions are a big business, as many newspapers, blogs, and broadcasts companies employ experts to make predictions of the outcome of the game. *The comparative accuracy of judgmental and model forecasts of American football games*, Song, Boulier and Stekler[4] used the predictions of 70 expert predictions of around 500 NFL football games over the course of two seasons and compared the accuracy ratio with statistical systems, and chance, to see if there is a significant difference between the two. The predictions analyzed were not just the simple "Wins" predictions, but they were also predictions made against the Las Vegas betting line. The findings were that there was no significant difference between the two. Even when you take into the factor of making predictions at the start of the season, and include predictions made near the middle of the season (at which point you can see clear front runners and struggling teams),

4. ChiUng Song, Bryan L. Boulier, and Herman O. Stekler, "The comparative accuracy of judgmental and model forecasts of American football games," *International Journal of Forecasting* 23, no. 3 (2007): 405–413, ISSN: 0169-2070, doi:https://doi.org/10.1016/j.ijforecast.2007.05.003, http://www.sciencedirect.com/science/article/pii/S0169207007000672.

the difference between the experts and the systems are not significant. In fact, it is just slightly better than chance.

Even with the rarity of turnovers in football (averaging less than 2 interceptions or fumbles per game in the 2019 season), Bock[5] attempted to create a prediction model using a Bernoulli distribution model function. He also used decision trees and other classification models, created in the programming language R, to see if it's possible to predict a percent chance of a turnover happening. Using collected data from 7 seasons, he saw that there was a confident way to predict the turnovers, especially putting into factor the previous play. If a team uses a predictable method of selecting the next play, and the defensive side can accurately predict whether the play will be a pass or run, the defence can force a fumble or interception much better.

Sports predictions are not limited to predicting wins or losses, but even which plays will be selected on the field. With *Using Open Source Logistic Regression Software to Classify Upcoming Play Type in the NFL*,[6] the article used 13 NFL seasons between rivals Pittsburgh Steelers and the Cleveland Browns. The authors attempted to predict the run or pass was attempted using data mining and logistic regression methods. The data was collected from a website and broken down to just 11 attributes for over 26,000 plays. The result created was approximately a 60% accuracy rate, which is low, but the authors mentioned promise that more attributes could help make the predictive model more accurate.

Machine learning is not the only way to examine the regression analysis of predicting the choosing between plays in the NFL. McGarrity and Linnen[7] discussed the decision between passing and rushing based on the difference between economic theory and game theory. They took 11 teams from the 2006 season where the starting quarterback got injured and the backup quarterback had to step in, and analyzed the difference in play calling between rushing and passing. According to the isoquant and isocost analysis, because the backup quarterback is less experienced, and therefore less reliable than the starter quarterback in passing, the offensive team will throw less. But after several tests of the coefficient between the two, the difference in the number of passing is not statistically

5. Joel R. Bock, "Empirical Prediction of Turnovers in NFL Football" [in English], Name - National Football League–NFL; Copyright - Copyright MDPI AG 2017; Last updated - 2017-08-23, *Sports* 5, no. 1 (2017): 1, `http://ezproxy.lib.ryerson.ca/login?url=https://search-proquest-com.ezproxy.lib.ryerson.ca/docview/1888935910?accountid=13631`.

6. Robert E. Baker and Ted Kwartler, "Sport Analytics: Using Open Source Logistic Regression Software to Classify Upcoming Play Type in the NFL" [in English], Name - National Football League–NFL; Cleveland Browns; Pittsburgh Steelers; Copyright - Copyright Sagamore Publishing LLC 2015; Document feature - Tables; ; Last updated - 2016-01-08; SubjectsTermNotLitGenreText - United States–US, *Journal of Applied Sport Management* 7, no. 2 (2015), `http://ezproxy.lib.ryerson.ca/login?url=https://search-proquest-com.ezproxy.lib.ryerson.ca/docview/1730027840?accountid=13631`.

7. Joseph P. McGarrity and Brian Linnen, "Pass or Run: An Empirical Test of the Matching Pennies Game Using Data from the National Football League," *Southern Economic Journal* 76, no. 3 (2010): 791–810, doi:`10.4284/sej.2010.76.3.791`, eprint: `https://onlinelibrary.wiley.com/doi/pdf/10.4284/sej.2010.76.3.791`, `https://onlinelibrary.wiley.com/doi/abs/10.4284/sej.2010.76.3.791`.

significant.

Predicting the odds of a binary action using regression is not limited to just football, as Healey[8] attempted to make a predictive model calculating the odds of a strikeout between the batter and the pitcher. Using a logit regression model called a Log5 model over a million matchups over a span of 10 seasons he was able to make a very accurate model to calculate the likelihood of a strikeout. He compared his results with the differences of left and right-handed batters and left and right handed pitchers. The big takeaway from this is that it's the batters that offer the biggest variance to the calculations.

With the popularity of the NFL draft and the evaluation of college players before the draft (known as the Scouting Combine, or just the Combine), many hope to find the next superstar to tear up the gridiron. Teramoto, Cross and Willick[9] took 9 years of the NFL combine data and attempted to determine the appropriate attributes that make a quality running back and wide receiver. Using multiple linear regression over the many tests done at the Combine, along with the player's physical attributes (like height and weight), the data was then compared to the first 3 years of their time in the NFL to see which attributes results of their physicals best reflect the quality of the player. The authors used Principal Component Analysis (PCA) to validate the data. In the end, how high they can jump can be a huge factor for wide receivers, and years played along with height and weight can make all the difference for running backs.

Predicting the outcome of a play is big when it comes to sports, but the most popular, and frankly the most difficult, is predicting the outcome of the game before it even starts. In *NBA Game Result Prediction Using Feature Analysis and Machine Learning*, Thabtah, Zhang and Abdelhamid[10] used multiple popular methods of prediction and regression, including neural networks, Naive Bayes and Decision Tree to predict the outcome of the games from the National Basketball Association (NBA). The amount of data that they accumulated for this test is over 35 years' worth, ranging from 1980 to 2017. Using both programming language R and the analysis program WEKA, and comparing all the results to each other, they were able to see a significance in attributes such as defensive rebounds, three-point percentages and free throws made the difference between winning and losing games.

With the National Hockey League (NHL) joining in on Big Data collecting soon, predicting win-loss games should be easier later down the road. In *A game-predicting expert system using big data and machine learning*,[11] the au-

8. Glenn Healey, "Modeling the Probability of a Strikeout for a Batter/Pitcher Matchup," *IEEE Transactions on Knowledge and Data Engineering* 27, no. 9 (2015): 2415–2423, doi:10.1109/TKDE.2015.2416735, http://resolver.scholarsportal.info/resolve/10414347/v27i0009/2415_mtpoasfabm.

9. Masaru Teramoto, Chad L. Cross, and Stuart Willick, "Predictive Value of National Football League Scouting Combine on Future Performance of Running Backs and Wide Receivers," *Journal of Strength and Conditioning Research* 30, no. 5 (2016): 1379–1390, doi:10.1519/JSC.0000000000001202.

10. Fadi Thabtah, Li Zhang, and Neda Abdelhamid, "NBA Game Result Prediction Using Feature Analysis and Machine Learning," *Annals of Data Science* 6, no. 1 (2019): 103–116, doi:10.1007/s40745-018-00189-x, https://doi.org/10.1007/s40745-018-00189-x.

11. Wei Gu et al., "A game-predicting expert system using big data and machine learning,"

thors attempt to make several predictive models using non-parametric statistical analysis, data mining, Support Vector Machine (SVM) analysis, and PCA to predict the outcome of NHL games. Because of the frequent rotation of players on the ice and the high speeds and action, predicting NHL game outcomes is much harder than all the other major sports. And with 90% accuracy, it seems that the SVM model was the most accurate out of all the predictive methods, with factors including the goalie's save percentage being the most important detail, as opposed to the skills of the star forward on the team, or home ice advantage.

Random forests seem to be among the most popular methods of creating a classification and regression model. For Schauberger and Groll's[12] article, they used random forest models to determine the outcome and scores of FIFA World cup scores based on factors like average age of the team, FIFA rank and betting odds. Although no prediction of future games was made, seeing how FIFA happens once every 4 years, using the covariate-based regression model seemed to bring in great results, especially testing out many methods of random forest; in particular, the Gamboost approach and the Lasso penalty.

Predictions in sports can also happen with Vegas odds. In the article *Are Sports Betting Markets Prediction Markets?: Evidence From a New Test*,[13] the authors used seemingly unrelated regression and the Breusch-Pagan test to determine if sports betting lines, and the over/under from sports gambling, can be used as a predictor. For betting lines, the objective is to maximize profits, and enticing a balance in bets so that winners get paid with the loser's money. Analyzing both the opening and closing prediction from Vegas on lots of games from the NFL, NBA and NCAA, and found that the over/under number is nowhere near a great way to predict the outcome of a game, but the betting line is an accurate predictor.

Data Set

In November of 2019, The NFL placed two data sets onto the website*Kaggle*[14]. Both data sets were part of two separate competitions with cash prizes and the response had a lot of great visualizations and machine learning model submissions.

Expert Systems with Applications 130 (2019): 293–305, ISSN: 0957-4174, doi:https://doi.org/10.1016/j.eswa.2019.04.025, http://www.sciencedirect.com/science/article/pii/S0957417419302556.

12. Gunther Schauberger and Andreas Groll, "Predicting matches in international football tournaments with random forests," *Statistical Modelling* 18, nos. 5-6 (2018): 460–482, doi:10.1177/1471082X18799934, eprint: https://doi.org/10.1177/1471082X18799934, https://doi.org/10.1177/1471082X18799934.

13. Kyle J. Kain and Trevon D. Logan, "Are Sports Betting Markets Prediction Markets?: Evidence From a New Test," *Journal of Sports Economics* 15, no. 1 (2014): 45–63, doi:10.1177/1527002512437744, eprint: https://doi.org/10.1177/1527002512437744, https://doi.org/10.1177/1527002512437744.

14. https://www.kaggle.com/c/nfl-big-data-bowl-2020

The data set used in this article is a collection of the number of running plays that occurred in the regular season of the NFL from week 1 of 2017 to week 12 of 2019. The data set provided included 49 attributes and 682,154 records. These records represent the 688 games played, and each game contains a total of 31,007 plays. Every play had 22 players on the field, 11 for the defense, and 11 for the offence, including the player running the ball.

The attributes that were provided are very detailed and shows that the NFL is serious about collecting and analyzing data. It included the teams in each game, the scores prior to the play, the down and distance needed, the yard line the play was on, and the number of yards gained from the play. There are also two columns that offer a nanosecond timestamp of when the ball was hiked, and when the ball was handed off to the running back. It also offered a lot of player information, including their birthday, height, weight, jersey number, and the university they played for. The stadium played was also involved, including whether the stadium is a dome or an outdoor field, the field surface itself, and several columns on the weather temperature. The play information was also in detail, including the formation of both the offense and defence, and even goes so far as to have the X and Y coordinates, speed and acceleration of the run, and the orientation and the angle of the player motion during the hike of the ball.

For the purposes of simplicity, the player performing the rush was extracted. The primary focus of the data consists of these 31,007 records. For the rest of the records, instead of just removing the data, it was normalized with the sum of each position. So, for example, if 3 of the 22 players on the field are defensive linemen, a "Linebacker" column populated a 3. Some of these columns will have low variance, which will be removed in the dimensionality reduction stage.

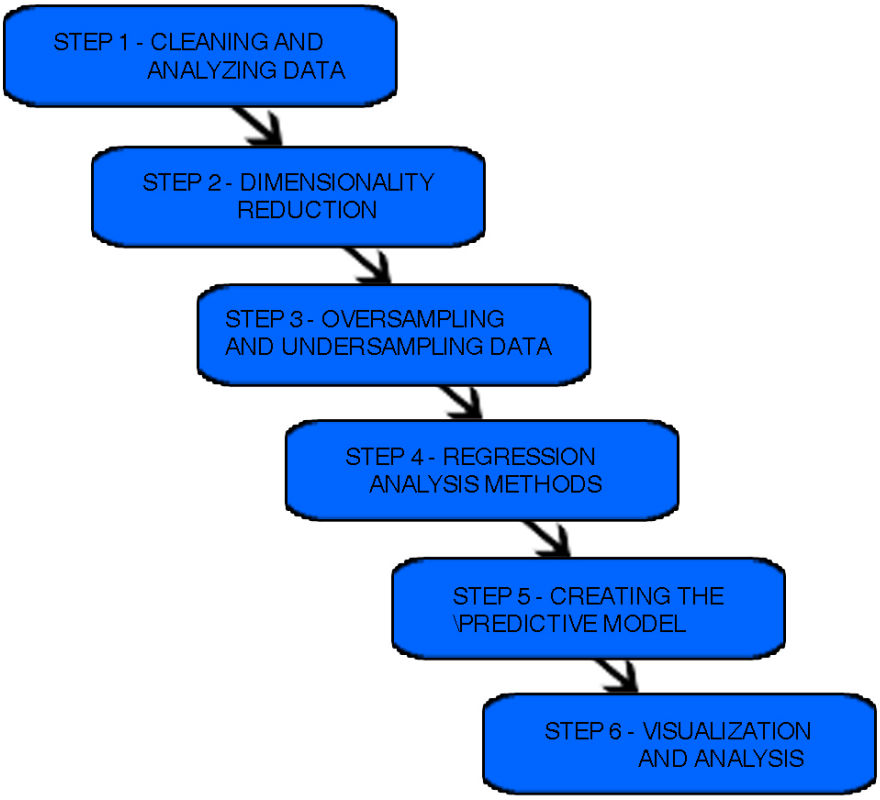
Below is the information provided by the National Football League on *Kaggle*. Many of these fields will be vital in making an accurate predictor model and regression analysis.

Table 1: Data dictionary provided by the data set

GameId	a unique game identifier
PlayId	a unique play identifier
Team	home or away
X	player position along the long axis of the field.
Y	player position along the short axis of the field.
S	speed in yards/second
A	acceleration in yardssecond$\hat{2}$
Dis	distance traveled from prior time point, in yards
Orientation	orientation of player (deg)
Dir	angle of player motion (deg)
NflId	a unique identifier of the player
DisplayName	player's name

JerseyNumber	jersey number
Season	year of the season
YardLine	the yard line of the line of scrimmage
Quarter	game quarter (1-5, 5 == overtime)
GameClock	time on the game clock
PossessionTeam	team with possession
Down	the down (1-4)
Distance	yards needed for a first down
FieldPosition	which side of the field the play is happening on
HomeScoreBeforePlay	home team score before play started
VisitorScoreBeforePlay	visitor team score before play started
NflIdRusher	the NflId of the rushing player
OffenseFormation	offense formation
OffensePersonnel	offensive team positional grouping
DefendersInTheBox	number of defenders lined up near the line of scrimmage, spanning the width of the offensive line
DefensePersonnel	defensive team positional grouping
PlayDirection	direction the play is headed
TimeHandoff	UTC time of the handoff
TimeSnap	UTC time of the snap
Yards	the yardage gained on the play (you are predicting this)
PlayerHeight	player height (ft-in)
PlayerWeight	player weight (lbs)
PlayerBirthDate	birth date (mm/dd/yyyy)
PlayerCollegeName	where the player attended college
Position	the player's position (the specific role on the field that they typically play)
HomeTeamAbbr	home team abbreviation
VisitorTeamAbbr	visitor team abbreviation
Week	week into the season
Stadium	stadium where the game is being played
Location	city where the game is being played
StadiumType	description of the stadium environment
Turf	description of the field surface
GameWeather	description of the game weather
Temperature	temperature (deg F)
Humidity	humidity
WindSpeed	wind speed in miles/hour
WindDirection	wind direction

Approach



Step 1 - Cleaning and analyzing data

The data was given as one single csv file, so no merging needed to be done. The original state of the data was collected electronically with the computer chips in each player's equipment thanks to the NFL's Next-Gen Stats. Because of this, there are very few situations with missing data. However, there were instances of data that needed to be corrected.

The biggest change to the data was creating the "First Down" attribute that is the new predictive variable. The provided variable was the number of yards that was gained from the field. Placing this attribute into a histogram shows a normally distributed collection of data, and it shows a mean of 4 yards and median of 3 yards, ranging from a 15-yard loss to a gain of 99 yards. So instead of using this as a predictor variable, a new binary field was created as to whether a first down was successful or not. A first down is calculated with the number of yards gained or lost, subtracted with the distance needed to gain the first down. Note that if the running play led to a touchdown, then it will count as a

successful first down gain for the purposes of this study.

The majority of the outliers found in the data were found to be a part of a natural variation. For example, Darren Sproles, running back for the Philadelphia Eagles, is 5 foot 6 inches, and is listed as an outlier in the "PlayerHeight" attribute. Since a lot of running backs vary in size, removing the records are very detrimental to the analysis. When creating a scatter plot with the player's height and weight, there were concerning dots found in the player's weight. However, because of a moderately high correlation coefficient, weight was removed instead. No records were removed due to the outliers found.

Because a lot of the stadiums had their own ways of inputting data, a lot of inconsistencies were found when adding the data. The primary example is the "weather" and "stadium type" fields. Weather had a text description of what the weather was like that day. The end result is using regular expressions to find the key words that will be populating noticeable trends that could affect running plays, like rain, wind, and snow. Stadium type had many different ways of saying whether the dome was open or not, so it had to be cleaned and narrowed down to 4 categories, outdoors, indoors, and domes with retractable roofs that were either open or closed that day.

The weather fields were not collected in several domed stadiums, like At&T stadium for the Dallas Cowboys and the State Farm Stadium for the Arizona Cardinals. Some of the temperatures can be fixed by filling in the average data. The other big problem was the data collecting for the stadium in England. For the few special games played there, the wind fields swapped in the wrong fields. They were corrected before any other changes were made. Several of the fields were text, like whether the possessing team was home or away, so a categorical number was created for each string field, with the exception of the Player information, which has the associated player id provided by the NFL.

The data provided a field position of where the ball was spotted, and with that, a field was added so that the yards needed for a touchdown was calculated. It's important to find out if the number of the success rate differs if the team is far away from scoring, or if the team is within 20 yards of scoring a touchdown (known as "the red zone"). One new field was added with subtracting the time of the hike and the time of the handoff. As expected, a lot of plays were within 1-2 seconds, but ranged from 0 seconds (the ball snapped straight to the runner) all the way to 7 seconds.

All cleaning and variable analysis will be performed using Python.

Step 2 - Dimensionality reduction

Due to a lot of fields added from the cleaning, even the ones that turned text into categorical numbers, we have a total of 66 attributes. For this, each step is going to be examined for reduction. Key fields that are removed first will be the text-based fields, because they have been converted to numerical. That action already removed 27 fields, concentrating only on the integer and float fields.

Several of the fields could be removed easily. The "Jersey number" will not be useful for analyzing because it's a categorical number. The "GameId", "PlayId",

and "PlayerId" are more identifiers, so removing them will not affect the analysis. The player's height was kept over the player's weight because the two fields are highly correlated with a coefficient number around 0.5, as mentioned earlier.

Removing the fields with very low variance and very high correlation was the next step. Several of the added fields containing a count of positions have very low variance, like the quarterback and halfback fields, so they were removed. There were also highly correlated position fields that we could narrow down to one, like the number of tight ends versus the number of wide receivers.

Random Forests, Principal Component Analysis, Forward Feature Construction and Backward Feature Elimination will occur during the assembling of the linear module, which more fields will be eliminated during analysis.

Step 3 - Oversampling and undersampling data

Because the number of successful first downs represent 21% of the total number of records, balancing the data will be essential for accuracy. Many methods will be tested for oversampling, including random oversampling and SMOTE. For undersampling, methods will include Tomek links and Cluster centroids will be used, as well as random undersampling. Decision trees in Python will be used to test each accuracy. Each method will be used for each model and the accuracy will be recorded.

Step 4 - Regression analysis methods

Using a Logistic Regression model in Python, using all forms of oversampling and undersampling will give us an insight as to which fields are the most important in determining the most important attributes to observe. Classification models, such as decision trees and random forest, will be assembled and examined to determine the popular factors. If the accuracy, precision and recall of the numbers turns out to be very low, creating logistic regression models in R will be done to verify the results. A ridge regression and a lasso regression will also be used in the hopes to find a more accurate model.

Step 5 - Creating the predictive model

Predictive modelling will be explored in order to determine if different methods can be used to interpret the data. A decision tree and random forest will be created using Python and R, both predictive and regression, which will give a visualization of the factors that greatly affect the running play. As well, a Naive Bayes model, Support Vector Machines, Neural Networks, and a K-Nearest Neighbors method will also be created and tested with Python and R. Throughout the processes, the accuracy will be collected, along with the precision, recall and f-score, and this will determine which model gave out the best result.

Step 6 - Visualization and Analysis

Once all the modelling is done, visualizations will be created in all the data, using MatPlotLib in Python, as well as tableau. Collecting all the accuracy percentages will help determine the best method of predicting the first down with all the data collected. The hope at the end is that we will have a good answer to which factors will be the most impactful and whether we have enough data to confidently say which running play scenarios are the best and which ones will lead to disastrous results.

References

Baker, Robert E., and Ted Kwartler. "Sport Analytics: Using Open Source Logistic Regression Software to Classify Upcoming Play Type in the NFL" [in English]. Name - National Football League–NFL; Cleveland Browns; Pittsburgh Steelers; Copyright - Copyright Sagamore Publishing LLC 2015; Document feature - Tables; ; Last updated - 2016-01-08; SubjectsTermNotLitGenreText - United States–US, *Journal of Applied Sport Management* 7, no. 2 (2015). http://ezproxy.lib.ryerson.ca/login?url=https://search-proquest-com.ezproxy.lib.ryerson.ca/docview/1730027840?accountid=13631.

Birnbaum, Phil. "A Guide to Sabermetric Research." https://sabr.org/sabermetrics/single-page.

Bock, Joel R. "Empirical Prediction of Turnovers in NFL Football" [in English]. Name - National Football League–NFL; Copyright - Copyright MDPI AG 2017; Last updated - 2017-08-23, *Sports* 5, no. 1 (2017): 1. http://ezproxy.lib.ryerson.ca/login?url=https://search-proquest-com.ezproxy.lib.ryerson.ca/docview/1888935910?accountid=13631.

Dastin, Jeffrey, and Anirban Paul. "Amazon, NFL reach $130 million streaming deal for Thursday night games: source," 2018. https://www.reuters.com/article/us-nfl-amazon-com/amazon-nfl-reach-130-million-streaming-deal-for-thursday-night-games-source-idUSKBN1HX3EP.

Gu, Wei, Krista Foster, Jennifer Shang, and Lirong Wei. "A game-predicting expert system using big data and machine learning." *Expert Systems with Applications* 130 (2019): 293–305. ISSN: 0957-4174. doi:https://doi.org/10.1016/j.eswa.2019.04.025. http://www.sciencedirect.com/science/article/pii/S0957417419302556.

Healey, Glenn. "Modeling the Probability of a Strikeout for a Batter/Pitcher Matchup." *IEEE Transactions on Knowledge and Data Engineering* 27, no. 9 (2015): 2415–2423. doi:10.1109/TKDE.2015.2416735. http://resolver.scholarsportal.info/resolve/10414347/v27i0009/2415_mtpoasfabm.

Kain, Kyle J., and Trevon D. Logan. "Are Sports Betting Markets Prediction Markets?: Evidence From a New Test." *Journal of Sports Economics* 15, no. 1 (2014): 45–63. doi:10.1177/1527002512437744. eprint: https://doi.org/10.1177/1527002512437744. https://doi.org/10.1177/1527002512437744.

McGarrity, Joseph P., and Brian Linnen. "Pass or Run: An Empirical Test of the Matching Pennies Game Using Data from the National Football League." *Southern Economic Journal* 76, no. 3 (2010): 791–810. doi:10.4284/sej.2010.76.3.791. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.4284/sej.2010.76.3.791. https://onlinelibrary.wiley.com/doi/abs/10.4284/sej.2010.76.3.791.

Ricky, Abhas. "How Data Analysis In Sports Is Changing The Game," 2019. https://www.forbes.com/sites/forbestechcouncil/2019/01/31/how-data-analysis-in-sports-is-changing-the-game/.

Schauberger, Gunther, and Andreas Groll. "Predicting matches in international football tournaments with random forests." *Statistical Modelling* 18, nos. 5-6 (2018): 460–482. doi:10.1177/1471082X18799934. eprint: https://doi.org/10.1177/1471082X18799934. https://doi.org/10.1177/1471082X18799934.

Song, ChiUng, Bryan L. Boulier, and Herman O. Stekler. "The comparative accuracy of judgmental and model forecasts of American football games." *International Journal of Forecasting* 23, no. 3 (2007): 405–413. ISSN: 0169-2070. doi:https://doi.org/10.1016/j.ijforecast.2007.05.003. http://www.sciencedirect.com/science/article/pii/S0169207007000672.

Teramoto, Masaru, Chad L. Cross, and Stuart Willick. "Predictive Value of National Football League Scouting Combine on Future Performance of Running Backs and Wide Receivers." *Journal of Strength and Conditioning Research* 30, no. 5 (2016): 1379–1390. doi:10.1519/JSC.0000000000001202.

Thabtah, Fadi, Li Zhang, and Neda Abdelhamid. "NBA Game Result Prediction Using Feature Analysis and Machine Learning." *Annals of Data Science* 6, no. 1 (2019): 103–116. doi:10.1007/s40745-018-00189-x. https://doi.org/10.1007/s40745-018-00189-x.