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*

Abstract

In American football, the objective of the running back is to be handed the football and to run as fast as he can, in the hopes that he can gain yardage. The long-term goal is to score a touchdown, and the short-term goal is to gain the first down. Recently, the National Football League (NFL) has been collecting data from the field and posted a dataset on Kaggle with a collection of running plays from the start of the 2017 season. This study investigates the provided data to identify the key factors needed using regression analysis and attempt to make a predictive model for the running back to successfully gain a first down. This was done using Random Forest, Decision Tree, a Logistic Regression and K-Nearest Neighbors machine learning methods in Python. What will be gained from this is whether we have sufficient information in the collected data to determine if teams can make a first down with the running plays they have in their current playbooks, and which factors can improve to for more successful running plays.

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

*Student, Ryerson University

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

game win streak. This story was popularized with the 2011 movie *Moneyball*, based on the 2003 book by the same name.

It's not just baseball that has seen an increase. In 2015, the National Football League (NFL) started collecting data for sports analysis by placing computer chips into the player's uniform. 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 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 or not we have sufficient information to 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 model 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, which earns the most potential points in a single play. 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 is 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 as 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.

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/.

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 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), 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

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.

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.

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 passes is not statistically 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 (such as 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

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.

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.

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 authors 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 (considering that 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

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," *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.

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 entice a balance in bets so that winners get paid with the loser's money. Analysis of both the opening and closing predictions from Vegas on lots of games from the NFL, NBA and NCAA, finds 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.

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. There are 17 weeks in a single season, and each team plays 16 games. 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 current 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 feature selection stage.

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

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

Variable Name	**Description**
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 yards/second^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)

Continued on next page

Table 1 – *Continued from previous page*

Variable Name	**Description**
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 provided as one single comma-separated (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, most of the data is well structured and didn't need much cleaning.

The cleaning of the data started with fixing one critical error that saw some data swapped in two columns. Once that was fixed, a new predictor variable was created from the one provided. The data given had a predictor variable to the number of yards gained, but one was created to determine if a First Down was successful by determining if the number of yards gained exceeded the number of yards needed to get the first down.

Because there were several fields that contained text, categorizing them was the next step. Then, after visualizing the data, the next step is to determine how many outliers are in the data, and whether records needed to be cleaned up before continuing. Checking and filling any missing data was the next essential step in the process of cleaning the data. To finish this step, the previously mentioned normalizing of the data to include the counts of player positions was completed, and the data was ready to be analyzed and ready to move on.

Step 2 – Feature Selection

With all the number of attributes in the data, the next crucial step is to eliminate more variables to a respectable level to make the machine learning easier. Because the machine learning methods in Python only work with integer and float data, removing all the text-based and time-based columns would be the first step. Next would be to remove the identifying variables because it shouldn't matter which player or which team is making the play. The study is only focusing on the play data.

Removing the fields with very low variance and very high correlation was the next step. Using the variance information for each feature removed some obvious columns, and a correlation heat map found more redundant attributes that can be cleaned up.

The next step in selecting the prime features are functions from the Sklearn package of Python which help with feature selection. A Variance Threshold, Backwards Elimination, Recursive Feature Elimination and the Embedded method were used onto the data, to pinpoint the fields that would optimize the machine learning process.

Step 3 – Experimental Design

Because the number of successful first downs represent 21% of the total number of records, balancing the data will be essential for the veracity of the tested

methods. The data was duplicated, and from that an oversampled dataset and an undersampled dataset was created to be fed into the machine learning methods, and the ways to measure the efficiency was collected. Each machine learning method used the same records to compare accuracy across the board. A third method will be run alongside the data with Principal Component Analysis (PCA). Before running the PCA data, it will be oversampled to make sure we have equal parts of successes and failures in the data.

Step 4 – Regression Analysis and Predictive Modelling

The most popular methods of regression and predictive modeling will be done to find the most important factors in finding the best results. Decision Trees and Random Forest algorithms will be created to find the features that impact a first down the most. A logistic regression model will be created to find the likely percentage odds of a first down being successful. And finally, a K-Nearest Neighbors was created with the data to find an accurate prediction.

All four methods will be using both the oversampled and undersampled data. As well, a PCA method will be run on all the methods. Twelve models in total will be created for this study. Each test will record the accuracy, precision, recall, the f1 score and the ROC-AUC curve score to see which method has the strongest prediction power.

Results

Cleaning the data was the most time-consuming, as it is essential that the raw data evolve into consistent data that can be analyzed and looked at. For the most part, the data was properly structured. At first glance, the data seemed to be collected automatically by computers, and little to no manual data seemed to be collected. There are still some issues that needed to be addressed.

One major error in the data needed to be corrected before any work needed to be done. Two fields, the wind speed and the wind direction, were in the wrong column for many of the records. During the coding to fix this, one major observation is found. A lot of the missing data and errors are specific for certain stadiums. For example, this error with the wind speed all came from the Wembley Stadium in England, when the NFL were doing international games to promote the league. And some of the missing data in the wind speed column came from stadiums like the U.S. Bank Stadium in Minnesota and Ford Field in Detroit, which does make sense because these stadiums are indoors. It's a minor issue, but it's recommended that the NFL address this to all the stadiums to keep the data consistent for collecting.

The biggest change to the data was establishing the "First Down" attribute that is the new response variable. The provided outcome 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. A new binary field was created as to whether obtaining a first down was successful or not. The new field is calculated with the number of yards gained or lost evaluated with the distance needed to achieve the first down. If the number of yards is greater than or equal to the distance needed, the record is flagged as a success. Note that if the running play led to a touchdown, then it will count as a successful first down earned for the purposes of this study.

Many of the outliers found in the data were found to be a part of a natural variation, so few were removed. For example, Darren Sproles[15], 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 would be 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, and height was kept.

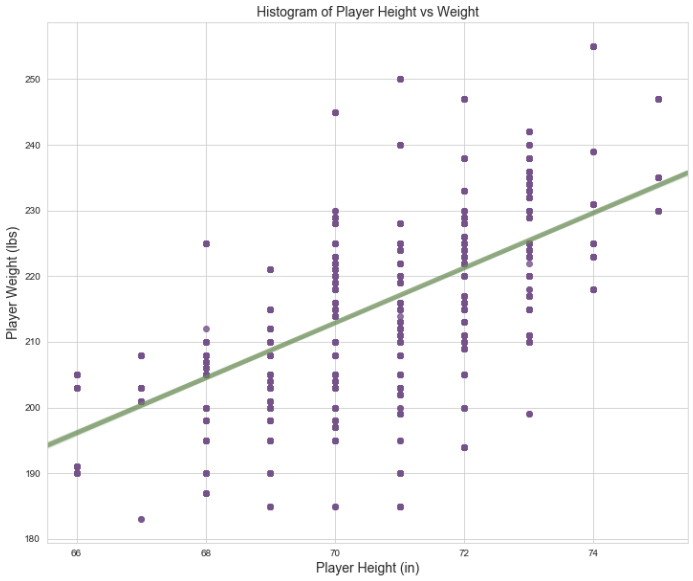

Figure 1: Correlation comparison between player height vs player weight

Three attributes had to be checked for outliers that were in contention for removal. The first one is the Quarter field. Normally a game has 4 quarters, but in case of overtime, an extra quarter is added to determine a winner in a sudden death format of the team who scores first wins. There are only 197 running plays that are in overtime in the entire data, and of those plays, 43

15. https://www.pro-football-reference.com/players/S/SproDa00.htm

were successful. These records were removed because it wouldn't affect the predictions in a negative way. The next one is the distance needed to make a first down. Most of the field is 10 yards, which makes sense since the start of drives and successful first downs start at first and 10. Keeping up to 20 yards would maintain a normal distribution. However, after 20 yards needed, there were 183 plays, and of those 1 was successful. These records can be removed, as these plays are extremely rare. We could discard anything above 10 yards due to rarity, but it would remove a lot more successes that we need to test the data. The final variable for outlier removal is the Position. For 93% of the data, the running back is the player performing the play. However, there are other positions that were collected. In recent years in the NFL, the quarterback has also made running plays, like Russell Wilson of the Seahawks and Cam Newton of the Panthers. However, for the purposes of this study, the focus will be on just the running backs. With all the outlier removals, the number of records is reduced to a still satisfying 28,694 records.

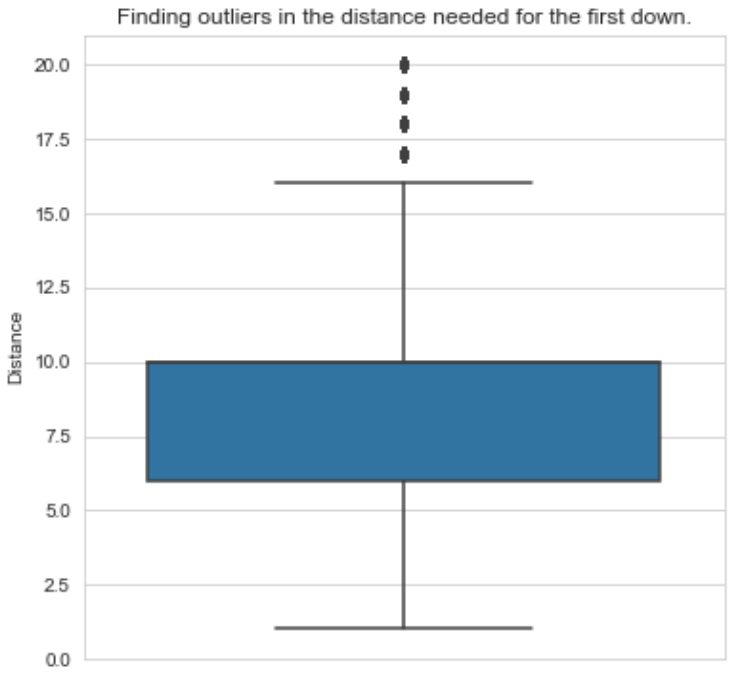

Figure 2: Outlier detection for the distance needed for a first down

Several of the fields were text, such as whether the possessing team was

home or away, so a categorical number was created for each string field except for 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 were calculated. It's possible 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 by 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.

In terms of the missing data, only 8 of the features had empty data that needed to be addressed. Several of the data fields were filled with the average data to keep the bell curve consistent. Several of the fields, including the Weather and the Humidity had more unknown cells that needed to be addressed. The attribute was considered again during the feature selection phase to see if it was worth keeping.

The final stages consisted of converting many of the text columns into a categorical numeric field. Fields like the team abbreviations had to be cleaned up before converting them into a standardized list. For example, Baltimore was abbreviated as "BAL" for the majority, but some listed it as "BLT", so it had to be corrected for all the teams with different abbreviations. Another cleaning was the Turf field. Turf was a list of products used, like the "FieldTurf 360" and "Twenty Four/Seven Turf", which are kinds of artificial turf. Because of this, each option was reduced to the fields of "Grass", "Turf" and "Hybrid", and then given a numeric option.

At the end of the cleaning phase, we ended up with 84 variables that will need to be reduced before any of the machine learning can be started.

Removing all the text-based columns and the timestamp fields is also necessary, since they are mostly redundant due to some of the column processing in the cleanup. Many machine learning programs in Python require integers and float numbers and don't allow strings, so it's safer to clean up them all so that we have consistent data for each method tested.

Removing identifier columns is the next step. Because teams like the New York Giants are far more successful at running plays compared to the Cincinnati Bengals, or that Alvin Kamara of the Saints has a more successful percentage of first down successes than Derrick Henry of the Titans, we want to remove the information that identifies the player or the team in the hopes to find a consistent answer.

The next part of the process is to look at the variance provided by the Python programming language. Here, we can see that a lot of the fields added during cleaning are already showing their value. In the columns that show the number of position players of the field we added from normalization, many of those show a lack of variation. To give an idea, there is a rarity to have more or less than 1 quarterback on the field, or 2 halfbacks, on their offense on the field.

The next step done was to create a correlation heat map, once done on

just the predictive variable, and one on the entire dataset. The idea is to find as many features that are like each other, and of those, one can be kept, and one can be tossed. In doing a correlation on just the first down variable, only one field was found to be above a threshold. The yards gained, which was the original predictor variable given, was correlated enough that we can justify removing the field.

When doing a correlation heat map on all the data, surprisingly there wasn't too many fields that met a threshold. For each combination of two correlated variables, a look at the data determined which field was expendable, and from that, several more fields were removed. So far, with all the steps taken, the number of attributes was reduced to 34.

The next step was to take some of the feature selection processes found on Python's popular Sklearn package to find more candidates worth removing. For each process, the data was checked and printed out the fields that were either worth removing, or worth keeping. No fields were removed on the following stage until each step was completed.

The first process up was Variance Threshold. In Variance Threshold, the function looks at the column's variance (as done in the step we did manually earlier) and eliminates the columns that don't meet the threshold. From running the test, 12 columns can be removed for having little impact.

The second step is the backward elimination process. Here, the performance of the data was checked with all the features, and one by one, the weaker features were removed until the model hit an acceptable performance. Using an Ordinary Least Squares method, 14 attributes can be removed, based on the p-value given by this run.

This was followed with Recursive Feature Elimination (RFE), which is similar, but uses the accuracy measurement to test the data. To test the number of variables that could be kept, a loop was created to run the RFE model onto a simple linear regression model, and the accuracy was collected in each iteration. Once the loop is completed, the best result is collected, and the RFE is run on the selected number to determine which attributes are the best. In this instance, the number of features it suggested to keep is 27, which is the best result so far.

The final stage is the Embedded Method. Using the lasso regularization, this method selects the attributes that contribute the most to the dataset and marks the one with no impact with a 0 coefficient. In this dataset, 22 variables were selected to be kept and 10 can be eliminated.

Post testing, the Recursive Feature Elimination method was chosen as the method to use. The machine learning models should have a lot of important variables to run the data on. During programming, it was found that the less attributes used in the machine learning methods, the less accurate the output became.

Once that was completed, and the data was split into oversampled and undersampled data, the 4 proposed methods of regression and predictive models were created. Each one had a third method of PCA run as well, for a total of 12 model run. Both oversampled and undersampled data was standardized before sending to the machine learning model scripts.

Table 2: Collection of quality measurements for all models

	Accuracy	Precision	Recall	F1	ROC	RMSE
Decision Tree, Oversampled	88.21%	83.22%	96.00%	89.15%	88.13%	0.12
K-Nearest, Oversampled	75.56%	76.30%	74.83%	75.56%	75.57%	0.24
Logistic regression, Undersampled	73.02%	76.09%	65.93%	70.65%	72.92%	0.27
Random Forest, Undersampled	72.91%	80.18%	59.76%	68.48%	72.71%	0.27
Logistic regression, Oversampled	72.43%	77.02%	64.69%	70.32%	72.51%	0.28
Random Forest, Oversampled	71.22%	73.37%	67.49%	70.30%	71.26%	0.29
Logistic regression, PCA	71.14%	74.04%	66.46%	70.04%	71.22%	0.29
K-Nearest, Undersampled	69.54%	73.90%	58.97%	65.60%	69.38%	0.30
Decision Tree, Undersampled	63.49%	62.82%	63.38%	63.10%	63.49%	0.37
Decision Tree, PCA	72.31%	34.16%	36.88%	35.47%	59.20%	0.28
K-Nearest, PCA	80.36%	56.45%	21.08%	30.69%	58.42%	0.20
Random Forest, PCA	79.38%	52.17%	0.68%	1.34%	50.26%	0.21

After running all four machine learning methods, the results were revealed. Each test collected a series of quality measurements, which are accuracy, precision, recall, F1 score, ROC score and the RMSE. From all the methods gathered, the Decision Tree on the oversampled data was the best result across the board. With an 88.21% accuracy and a ROC score of 88.13%, the data surpassed all expectations. The next highest was the K-Nearest Neighbors model with Oversampled data, with a 75.56% accuracy. The difference between the two have a large gap in-between each statistic, even in the RMSE, with Decision Tree getting a 0.12 and K-Nearest getting a 0.24. The best performing model using undersampled data was the Logistic Regression model with a 73.02% accuracy. The worst performing data came from the PCA, going so far as getting a 1.34% F1 score in Random Forest.

Table 3: Collection of quality measurements for all models

Feature	**Importance**
Distance	24.80%
A	11.00%
S	10.80%
Y	8.60%
YardsToTouchdown	7.10%

Collecting the feature importance from the Decision Tree Oversampled data gave us a better idea of which variables affected the prediction the most. Of all the columns given, the "Distance" column impacted the most. This showcases that, how far a player was from getting the first down impacted how successful they were at getting a first down. As mentioned before, the average number of yards gained is 4, and the median yards gained is 3. When broken down per team, the number of yards gained on a successful first down compared to the number of yards gained on a failed first down is strikingly similar. Each team averaged 2 yards gained on a failed first down, and the number of yards gained on a successful first down was between 8 and 12 yards, Even compared to which down the play was on gave interesting details. On first down, the odds of a successful run to get a first down was only 12%. This percentage increased with each subsequent down.

The second and third most important features from the Decision Tree is the

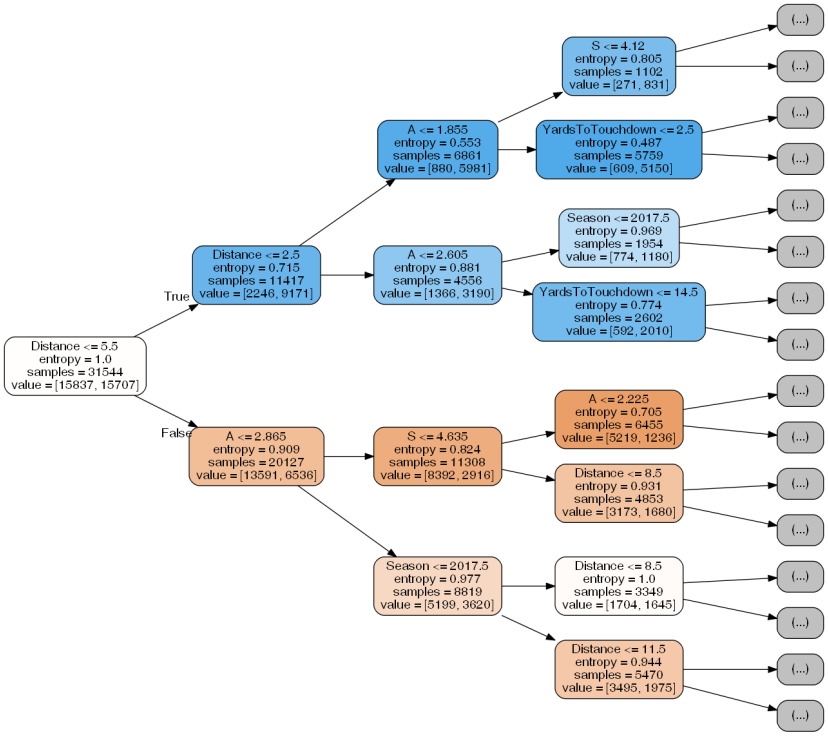

Figure 3: Results of the Decision Tree Oversampled Data

Speed and the Acceleration of the running back. So if the player is very fast, the odds of them gaining the first down increases as well. This is a very popular assumption in popular football circles already. The idea of the running back is to quickly run past any defending player by any means necessary to gain yardage. Normally, a running play will run dead in its tracks when a defensive lineman is ready to tackle the running back before they pass the line of scrimmage, and when that's successful, the number of yards gained is minimal, and could even go backwards in certain scenarios. But with a fast running back, and with speed and agility to bob and weave around any defenders, more yards are gained.

To look further into this, the data was looked at again, and this time taking all the players that had at least 400 carries in the data and calculated the percentage of successful first downs compared to the number of attempts. From these, a top 10 list was created of the most successful running backs. Some of the names on that list are some of the more popular players, including last year's MVP Derrick Henry of the Titans, Todd Gurley of the Rams, and topping the list is Alvin Kamara of the Saints. Funnily enough, the top 5 players with the best percentage are all 26 years old or younger, and with a 40-yard combine of between 4.50 and 4.58, which is an amazing time for a running back. For context, the 40-yard combine is a running test done before the NFL draft so

scouts can watch and analyze players before selecting them for their respective teams. Each player does a 40-yard dash and their time is recorded. The average 40 time for all running backs is 4.49[16]. This means that the top successful players can run fast, and gain speed quickly.

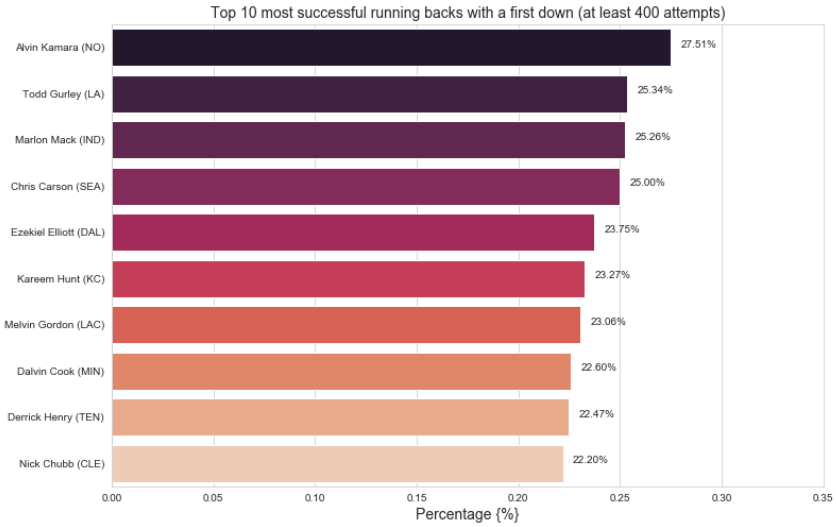

Figure 4: Top 10 most successful running backs based on percentage of first downs

The biggest takeaway from all the tests done is that the Decision Tree with Oversampled data performed well. In fact, it performed too well. Looking at the measurement numbers, one of the big fears of what came from the project is the fear of overfitting. Overfitting occurs when the incoming data in a machine learning model has too much unnecessary information, or noise. With the volume of attributes and records in the dataset, the potential amount of noise may be an issue.

As mentioned, the accuracy of the Decision Tree came up with an 88.21%. In contrast the best performing Random Forest in terms of accuracy is with the undersampled data with 72.91%. Normally the Random Forest does better than the Decision Tree because the Random Forest finds the best possible decision tree. Some of the possible factors is that the number of allowed outliers may have done more harm than initially thought. Many were kept in the hopes that because a lot of them fall under a natural progression, they should be safe to use. To try to examine it further, some experimental data manipulation was happening to the side, not affecting the current models. This experimenting

16. `https://www.milehighreport.com/2013/2/12/3969128/some-clarification-is-in-order-average-speed-by-position`

was working with a copy of the data, attempting to see if more outliers could be cleaned up for better numbers. In the end, the measurements didn't change enough, and the data was left how it is currently described. One method to try for down the field is using a different method of experimental design, like k-fold cross-validation. This method of data splitting could find a more optimum set of data to test the models with and provide a better story on that the data is presenting.

Another key reason for the potential overfitting is the simple fact that there still isn't enough data. The entire dataset represents just under three seasons of running plays. The NFL started collecting data in 2015, which means this data is still in its infancy. Much like many of the articles previously mentioned in the literature review, this could mean that more data needs to be collected to see if the current assumptions are correct. But in terms of what was found currently, there is hope that the more data we collect, the more accurate our models could become.

Conclusion

The National Football League has given the Kaggle community a a good collection of data to evaluate and analyze. In the end, the Decision Tree with oversampled data performed the best, and showed that the key fields that have the most impact are the distance needed to gain the first down, along with the running back's speed and acceleration at the moment of the ball snapping. However, since the data is relatively small, being only from less than three seasons, and with such high measurement numbers, the potential for overfitting needs to be acknowledged with accepting these results. That being said, this data is still very beneficial to not just the curious armchair quarterbacks looking for the factors needed to select for their next fantasy football team, but also very big for team coordinators and scouts who need to find the next big running back and assemble the plays that could bring in more successful plays. Because the more first downs you gain, the more you have the ball, and the more touchdowns you can gain to win games and even win championships.

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.