NFL Pass, Rush, and Receive Statistics: Analysis Based on Winning Strategies

Course: ADTA 5940 Advanced Data Analytics Capstone Experience

Team Members:

Ravi kannegundla

Sai Meghana Boyapati

Sai Kumar Reddy Yenumula

Naveen Challagundla.

Date: 03-20-2023

Introduction

Football League

The foundation of the National Football league was made in 1920 with the name of the American Football Association also termed APFA. In this league, there are around ten teams from four different states at the time of the beginning season. The main reason to build the NFL was to give a proper and organized professional league (NFL history, 2022).

As time goes on, football has changed its name from AFPA to NFL, and also widen its target market. In the meantime, the NFL has made itself the professional football league of the United States. The NFL first introduced its first championship, which was then known as “Super Bowl. (Super Bowl, 2022)”

The time between 1940 to 1950 was the time of the establishment of the NFL with the cultural force by getting amazing players like Jim Brown, Johnny Unites, and the Bart Star. The players attracted audiences around the country. The NFL kept expanding its market by adding the states of Loss Angles, San Francisco, and Baltimore (Peterson, 2017).

The NFL exponentially increased its growth by adding more teams in 1960 and introducing the Super Bowl as a major cultural event. During this time the league witnessed new star players like Joe Namath and Frank Tarkington (Peterson, 2017), (Super Bowl History, n.d.)

The NFL keeps growing by making the combination with AFL in 1970. The AFL is also known as the American Football league which was operated in the United States between 1960 to 1969. At this time new rules were introduced to make sure the safety of the players, in this league the first quarterbacks and coaches were seen. The AFL

was built by a group of a businessman who saw the potential to compete with the NFL, the AFL was founded with six teams (NFL history, 2022).

In the analysis of the NFL team, it is observed that NFL is facing some challenges concerning the safety of the players, though the league is constant to be the cultural force in America. Million people have started to watch the game and the Super Bowl is the major body to host the event each year.

Super Bowl

History shows that the first Super Bowl was played on the 15th of January, 1967 between the NFL's Green Bay Packers and the AFL Kanas City Chiefs. In 1970 the Super Bowl has become a good cultural event, with a halftime show which features important musical acts and booming love in the commercials which was displayed during the game. The aired commercial during the game, aided the game to become the major platform for the advertisement (Super Bowl History, n.d.).

Recently, the Super Bowl has got the position of the most-watched TV event in the world and become a major part of American culture. mostly, the game is used as a platform for political and social messages. The Super Bowl is known major revenue-generating body for both NFL and the host city.

Super Bowl LVI

The Super Bowl LVI the 56th edition is also the championship game of the National Football league. The game was played on the date of 13 February 2022. The game was played between Ram Angles and Cincinnati Bengals. The game was played at the Sofi Stadium in California Inglewood and the referee was Ronald Torbert. The Rams won the game by 3.5 (Super Bowl LVI, n.d.), (About the Stadium, n.d.)

The game is going to feature the NFL championship by displaying the champions of the American Football Conference (AFC) and the National Football Conference NFC,

it is anticipated that the game will get the attention of millions of viewers around the world. the broadcast of the game will be held in the USA and displayed on NBC and the website and its application (Super Bowl LVI, n.d.).

The terminologies related to the NFL are learned from (Eisenberg, 1920), (Bissinger, 1970):

Quarterback: it is the player who directs the offense and throws the ball. In American Football offense word is used for the team who had a ball that throws the ball to the opposing team. The offense got the four attempts which are also known as downs, in these attempts the offensive team is supposed to throw the ball 10 yards down the field. In any case, if the team failed to cover 10 yards, the ball possession is the return to the opposing team, and if they succeed, the team is given another four attempts.

Running back: it is the player who holds the ball on running plays. In NFL, running play is a kind of offensive play where the ball is passed to the running back and the person tries to advance by running forward. The purpose of the running back is to find the gap in the defensive line and cover as many yards as possible before the opposite team defender’s tackle.

Wide receiver: it is the player who catches the passes of the quarterback. The wide receiver is one of the types of the offensive player who has the potential to catch the passes. The wide receivers typically position to the edges of the offensive structure to increase the player's ability to run downfield and catch long passes.

Tight end: he is in the position of the offensive team. The tight-end player has both powers to receive and block the ball. The position of the tight player is assigned as, he is either next to the offensive tack or separated like a wide receiver, it is named like that because the position to the player is assigned to be tightly next to the offensive line.

Offensive line. These are the group of players who blocks to get the quarterback and run back.

Defensive line: these are also the bunch of players who try to block the offense from advancing.

Linebacker: it is the defensive position that is lined up behind the defensive lineman and in front of the secondary. The linebackers are divided into three categories, middle linebacker stops the opposite side team running game in addition to that performs the task to cover receivers in passing situations. The strongside linebacker is lined up to the side of the offense where the tight end is lined up. However, a weakside linebacker is lined up to the side of the offense where is not any tight end.

Defensive back: it is the position on the defensive team. There are groups of players who are assigned the task to cover receivers and try to stop the offense from scoring.

Data Analysis

The data analysis is used to process the given data to get the information, and insights and conclude on that basis. The data analysis technique is important for deciding based on available data. the data analysis technique is used in the fields like business, health, social sciences, technology, and engineering. The recent development of technologies has made data analytics the hot field for all sectors and its use is increasing the amount of data is increasing day by day.

Another step in data analysis is the exploration and visualization of the data. the data is visualized by using different steps such as histograms, scatter plots, and heat maps. The data visualization techniques help the analysts to learn about the relationships and patterns in the dataset.

The predictions, hypotheses tests, and identification of significant differences are made on the bases of statistical analysis. The final step of data analysis is the

interpretation based on the statistical analysis. in the interpretation, the data analyst presents the results clearly and concisely and makes the recommendation based on that analysis.

Research Questions Analysis

In this paper (Chen, 2019), the author focuses on the long passes and pass rating impact on the chances of winning the football game. The author gave a statistical analysis for checking the impact. Football has different factors. The author used the data of long passes and pass ratings from the 200 NFL games which were played in 2018-2019 and performed the regression analysis to check the relation between long passes and the pass ratings and their impact on game-winning probability. After checking the author conclude that both long passes and pass ratings have a significant positive impact on game-winning chances. The author proposed that for every long pass, the chances of winning the game incremented 3% and with the addition of every pass rating point, the winning chances increased by 6%. Here, the author shows the percentage of a rating point hired, so the team must focus on the pass ratings to get a higher chance of winning. It is no doubt that the paper has provided some useful insights on long passes and pass ratings winning chances in the football game but there are some limitations, the first limitation is that author used the dataset from 2018-2019. The size of the sample from the dataset is reasonable but the results could not be generalized to other seasons or leagues. If the data is used from various leagues and seasons, this will increase the generalization of the results. the other limitation is that the research of the author focused on two performance factors that impact the winning chances of the football game. The last the author used regression analysis, which uses the statistical technique, but it does not establish causality.

In different football skills, passing also plays a major role in the success of the team, the paper (Barnes, 2016) investigated data from the 2012 NFL and learned that the percentage of completion has an impact on the game-winning chances, the author proposed that the team having higher completion are more effective in moving the ball down in the field which results in more score and more chances of winning. Another paper (Krzyzewski, 2018) used the data from the 2014-2016 NFL season to investigate the relational graph of completion percentage and the success of the team. The author concludes that the higher completion percentage lead to better offensive efficiency and there are chances to win the game. The author also found that the parameter of completion percentage shows better performances and some other factors such as the yards per attempt. In short, the research of the two authors proposed that completing passes is the major factor in the winning chances of the football game. The high completion is like better efficiency, doing more score chances, and better control of the pace of the game. The papers had limitations in the data used like 2012 NFL data and 2014-2016 NFL seasons, which could not be used as representative of other leagues another limitation is the authors did not focus on other factors such as yards per attempt, turnover, and penalties.

In this paper, the author analyzed the home advantage in European 10 different countries. The author analyzed data from 15 thousand matches from the year 2011-2016 seasons. The author applied a logistic regression model to investigate the relationship between the probability of a home win and different other factors such as gender, score, and the league. The author concludes that home advantage varies by country and league, in some teams it showed stronger home advantage, and in some the weaker. the Spanish Liga had the highest winning percentage and the Dutch Eredivisie showed the lowest percentage in terms of home wins. The author

concluded that the home advantage is higher for women than men in football. However, the home advantage has decreased in recent years which is due to the amendment in the rules and regulations. The introduction of the goal line technology has decreased the number of wrong decisions by the referee which has reduced the home advantage.

In this paper (Brown, 2019), the author investigates all two points attempts in the NFL from the year 2001-2017, and realized that the teams are encouraged to attempt two-point conversions when they are on the verge of losing the game, having less time remaining and are playing against the stronger team. the successful two-point conversion had the higher chances of winning the game. The author also noted that the coaches prefer to attempt two-point conversions when losing by fourteen points than when losing 15 points (as need three scores). Another paper (Madden, 2016) also used statistical modeling to investigate the strategies in the decision-making of NFL coaches regarding two-point conversions. The success rate of the attempts increases the point differential and the time residual in the game. here, the author found that coaches are overly conservative if they had to attempt two-point conversions. The author noted that the teams are trying to attempt two-point conversions usually when they do, particularly when they have strong offenses. It shows that the team would have the advantage to be more aggressive in attempting two points conversions.

Data analysis Case Studies Review

The paper (Kirkos, 2007) is about the use of data mining techniques in the detection of fraud activities in financial statements. The paper explains the effectiveness of the data mining classification in the detection of fraudulent activities in financial activities. Machine learning techniques like decision trees, neural networks, and Bayesian belief networks are used in the identification of malware in financial statements. The paper used three models to compare the performance. the three methods used are decision trees, neural networks, and Bayesian belief networks. The results of all three are given below:

Model	Fraud (%)	Non-Fraud (%)	Total (%)
Decision Tree	92.1	100.0	96.2
Neural Network	100.0	100.0	100.0
Bayesian Belief Network	97.4	92.1	94.7

Table 1 results of all three models

In this table, the models are applied to the training sample. The table shows that the neural network shows the best performance and is quite efficient in separating the fraudulent and non-fraudulent activities, after the neural networks Bayesian belief networks show the better performance. The models are validated to make sure there is no biasness by ordering the training data. Here, the author finds out that in some cases the models just memorize the sample from data other than learning that concept is known as “data overfitting”. The author chose a 10-fold cross-validation approach. In this process, the sample is distributed in 10 folds. The author made sure that each fold has the same number of fraudulent and non-fraudulent activities. the model is trained

by using the remaining nine folds and being tested by the means of holding out the fold. In the last, the author calculated the average performance.

The following table is presented by the author after cross-validation

Model	Fraud (%)	Non-Fraud (%)	Total (%)
Decision Tree	75.0	72.5	73.6
Neural Network	82.5	77.5	80.0
Bayesian Belief Network	91.7	88.9	90.3

Table 2 Model results after 10-fold validation

As anticipated that accuracy of each model is changed after the validation of the dataset. Each model shows a considerably different performance. Here the Bayesian Belief Network performed the best and the neural network second in the performance.

Student Academic Performance Analysis

In this paper (Kim, 2020) machine learning algorithms are used to predict the performance of the students. The paper focuses on each student's academic performance with the help of machine learning algorithms. The author gathered data from the Korean university Learning Management System (LMS). The content is the course grades, the record the attendance, the score of the assignments, and the information from the demographics. Machine learning algorithms such as logistic regression, decision trees, random forest, and gradient boosting were used. the objective is to study the efficiency of the machine learning algorithm for the prediction of student academic performance. the total number of students in the dataset is 4763 students from the Korean university. The random forest produces the best result with an accuracy of 87.9%. the other three algorithms' results start from

the logistic regression which has the lowest accuracy, the accuracy of logistic regression is 69.6%. the accuracy of the decision tree is 74.2% and the accuracy of the gradient boosting machine learning algorithm is 83.6% which is the second highest. After the implementation of the machine learning algorithm, the author comes to know that the different factors which significantly impact the academic performance of the student are the type of admission, gender, and attendance records.

The paper (Wiyono, 2020) explains that several studies have been done so far for the improvement of the academic performance of the students, the methods like data mining algorithms which were used to do the student performance analysis, analyze the student performance using the clustering techniques to predict the student performance on the base of (poor, average, good and excellent) using the provided dataset. In this paper, three machine learning methods are used for model evolution. The machine learning algorithms are K- nearest neighbor (KNN), support vector machine (SVM), and Decision Tree. The data is split into 75 percent for training and 25 percent for testing. The data used for the model is the training data, the samples uses for training data are 1148 samples, 6 predictors, and two classes. The cross-validation is the same 10 folds for this as well. The best result is produced with k nearest neighbor with k=3 (kernel). The accuracy for that model is 94.5% and the value of C=1 for the support vector machine with an accuracy of 95.09%, and cp=0.67 for the decision tree algorithm, and the result of the decision tree algorithm on that value is 95.65%.

The table for the model accuracies is given below:

KNN (%)	SVM (%)	Decision Tree (%)
94.5	95	93

Table 3 Accuracies of Supervised classifiers

It is shown that the best accuracy is produced by the support vector machine (SVM) machine learning algorithm, after that the accuracy of KNN is better in the prediction of the student performance.

Machine Learning Models for Sport Analysis

This paper will present the prediction approach based on team performance wot the data envelopment analysis method and also the data-driven technique used. the approach is divided into two steps: the first one is through the multivariate logistic regression analysis to check the relationship between the winning probability and the outcomes of the game. The other is the DEA methodology-based player profile (Li, 2021).

This paper (Landers, 2018) solves two related problems in the field of fantasy football in the Nation Football League (NFL). The author's approaches included using a machine learning algorithm to predict the points of fantasy players. These predictions will be used for the construction of optimal teams. The data from the FanDuel of the 2016 season from the National Football League is used to carry out the investigation. The objective of the author was to choose a team having one quarterback QB, two running backs, three wide receivers, one tight end, one defensive player, and one kicker. the individual player scores fantasy points on the of his performance in the real game. The goal of the author is to divide it into two steps Regression and Optimization. In the regression, each player's fantasy points in NFL before the game are predicted. For the optimization, the prediction points are used to develop a team with the highest possible number of points satisfying the total salary constraint. The dataset used from FanDuel has 256 games of thirty-two teams in the last seventeen weeks of all seasons of the 2016 NFL. In the experiment, the data is split into 11 folds

for cross-validation, each set for testing the one week of the 2016 NFL season. The author has noted that the big nature scale of fantasy sports gives a perfect setting for machine learning and artificial intelligence research. The problems had similarities to the research on multi-agents, cohesive formation, and general sports artificial intelligence.

The paper (Alonso, 2022) analyzes basketball's previous history match data. the author suggests that the prediction of the result depends on different factors such as coaching strategy, the skill of the player, and the morale of the team. Due to a large number of factors, it is difficult to get the exact result of an individual match. The paper focuses on learning about the basketball's previous match dataset. The features in this learning are included both individual and team which are used to predict the upcoming match. The study author focuses on the case study of the National Basketball Association (NBA). The author has carried out a comparative analysis to find the algorithm which gives the best prediction results. the author has used the classifiers such as k-nearest neighbor (KNN), the decision tree, and the Naive Bayes classifier. The author has used the Kaggle dataset, the advantage of using this dataset is that it does not contain any missing values. However, the dataset is huge with 4920 rows and 61 columns. The NBA paper (Al-Jarrah, 2015) is based on the use of supervised machine learning for the analysis of NBA basketball matches. The KNN is implemented on both non-reduced as well as PCA decomposition-obtained datasets. after KNN, the other more Naive Bayes and decision tree were implemented. The naive Bayes is trained with non-reduced data. the accuracy of 70.5%, which might not be the highest, but it made sure the balance of accuracy and recall was with the best proportions. The KNN also gives a good result after feeding it with a reduced PCA dataset. it gives 70% accuracy, uniform between the recall lost classification.

advantage of using this classifier is that it gives quick execution, which is lower than another version of it. In the last, the decision tree could not exceed 60% accuracy, and the balance of recall and its attribute was not quite impressive, to the author the reasons for the low results could be that dataset is not huge enough and the multicollinearity of the dataset features where were get confused by the decision tree classifier.

References

About the Stadium. (n.d.). Retrieved from SoFi Stadium.com: https://www.sofistadium.com/about-the-stadium/

Al-Jarrah, O. Y. (2015). Efficient machine learning for big data: A review. *Big Data Research 2.3*, 87-93.

Alonso, R. P. (2022). Machine learning approach to predicting a basketball game outcome. *International Journal of Data Science 7.1*, 60-77.

Barnes, C. H. (2016). The impact of game location and final score differential on game-related statistics in professional rugby union. *International Journal of Sports Science & Coaching*, 11(5), 684-689.

Bissinger, H. (1970). Friday Night Lights: A Town, a Team, and a Dream.

Brown, R. M. (2019). The decision to attempt a two-point conversion in the National Football League: A case study. *International Journal of Sport Finance*, 14(4), 296-314.

Chen, Y. L. (2019). The impact of long passes and pass ratings on the probability of winning a football game. *Journal of Quantitative Analysis in Sports*, 15(3), 153-163.

Eisenberg, J. (1920). The League: How Five Rivals Created the NFL and Launched a Sports Empire.

Kim, M. K. (2020). Predicting student academic performance using machine learning algorithms: A case study of a Korean university. *Computer and Education, 144, 103713*.

Kirkos, E. C. (2007). Data mining techniques for the detection of fraudulent financial statements. *Expert systems with applications*, 995-1003.

Krzyzewski, M. S. (2018). Analysis of passing efficiency and success rates in the National Football League. *Journal of Quantitative Analysis in Sports*, 14(1), 1-12.

Landers, J. R. (2018). Machine learning approaches to competing in fantasy leagues for the NFL. *IEEE Transactions on Games*, 159-172.

Li, Y. L. (2021). A data-driven prediction approach for sports team performance and its application to National Basketball Association. *Omega 98 102123.*

Madden, C. S. (2016). Two-point conversion strategy in the NFL: A quantitative analysis. *Journal of Quantitative Analysis in Sports*, 12(2), 43-56.

NFL history. (2022). Retrieved from NFL.com: https://www.nfl.com/history/

Peterson, R. (2017). The NFL: A history of America's favorite sport. Minneapolis, MN. *Lerner Publishing Group*.

Super Bowl. (2022). Retrieved from Wikipedia.com: https://en.wikipedia.org/wiki/Super_Bowl

Super Bowl History. (n.d.). Retrieved from NFL.com: https://www.nfl.com/super-bowl/history/

Super Bowl LVI. (n.d.). Retrieved from NFL.com: https://www.nfl.com/super-bowl/

Wiyono, S. e. (2020). Comparative study of KNN, SVM, and decision tree algorithm for student's performance prediction. IJCSAM) International Journal of Computing Science and Applied Mathematics 6.2 (2020): 50-53.