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

Course: ADTA 5940 - Analytics Capstone Experience

Team Members:

Ravi Kannegundla

Sai Meghana Boyapati

Sai Kumar Reddy Yenumula

Naveen Challagundla.

Date: 05-08-2023

Contents

Chapter 1: Introduction

Background Information

This paper discusses secondary data on the National Football League (NFL), a professional game popular among Americans. The secondary data being analysed for this analysis is based on a football league comprising thirty-two clubs split evenly between the American Football Conference (AFC) and the National Football Conference (NFC). In addition, the NFL is played here in the United States and Canada and is rated as one of the highest professional-level games in America. It is also recognized globally.

The NFL provides a wealth of data for advanced sports analytics through its official data provider, Sports Radar. This data includes play-by-play information, player statistics, and game-related metrics (NFL Raw Data Download, n.d.). Additionally, other sources of NFL data are available, including websites that provide analysis and visualization of the data, such as Pro Football Reference and NFL.com. The most used data sources for advanced analytics in the NFL include.

Our Research Analysis on the NFL will proceed as follows; We will outline our Research Question, state our hypotheses, describe the intended data set, and operationalize our variables for this analysis. Finally, we want to find a model that best fits our analysis.

Research Questions

1. Can the probability of winning a game affect a team's chance by long passes and/or pass ratings?

H: Long passes and higher pass ratings have a positive impact on a team's probability of winning a game.

2. To what extent does completing passes affect the likelihood of winning a game?

H: Completing passes has a significant positive impact on a team's likelihood of winning a game.

3. Is the probability of winning a game highly likely for the home team?

H: The home team has a statistically significant advantage in terms of winning games compared to the visiting team.

4. Does two-point conversions have the potential to contribute to a team's victory in a game?

H: Successful two-point conversions have a positive impact on a team's likelihood of winning a game.

Chapter 2: Literature Review

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 statistical analysis for checking the impact was carried out. It is obvious that football has different factors. The data was used for long passes and pass ratings from the 200 NFL games played in 2019-2023. The regression analysis was performed to check the relation between long passes and the pass ratings and their impact on game-winning probability. After checking, the author concluded that both long passes and pass ratings have a significant positive impact on game-winning chances. 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, it is observed that the percentage of rating points is higher, 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 2019-2023. 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 regression analysis was used; it uses the statistical technique, but it does not establish causality.

In different skills in football, 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 a higher completion percentage lead to better offensive efficiency, and there are chances to win the game and also found that the parameter of completion percentage shows better performances in 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 should have focused on other factors such as yards per attempt, turnover, and penalties.

In this paper, the author did the analysis on the home advantage in 10 different European countries and analysed 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. Also concluded that home advantage varies by country and league; in some teams, it shows a 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 referees that have 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 realizes 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 modelling 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 must 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 of being more aggressive in attempting two points conversion.

Research Survey

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.

In the paper, 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 fraudulent and non-

fraudulent activities; after the neural networks, Bayesian belief networks show better performance. The models are validated to make sure there is no business by ordering the training data. Here, the author finds out that in some cases, the models 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 means of holding out the fold. In the last, the author calculated the average performance. The results after the cross-validation are improved.

As anticipated, the 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 the 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 author gathered data from the Korean University Learning Management System (LMS). The content is the course grades, the record attendance, the score of the assignments, and the information from the demographics. Machine learning algorithms such as logistic regression, decision trees, random forests, 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, analyse 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 the model evolution. The machine learning algorithms are K-nearest neighbour (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 used for training data are 1148 samples, six predictors, and two classes. The cross-validation is the same ten folds for this as well. The best result is produced with k nearest neighbour 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%.

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 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 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 sport artificial intelligence.

The paper (Alonso, 2022) analyses 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 many factors, it isn't easy 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 of the 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 neighbour (KNN), the decision tree, and the naïve 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 68 columns. The NBA paper (Al-Jarrah, 2015) is based on the use of supervised machine learning to analyse NBA basketball matches. The KNN is implemented on both non-reduced as well as PCA decomposition-obtained datasets.

After KNN, the other naiver Bayes and decision trees were implemented. The naïve Bayes is trained with non-reduced data. The accuracy of 70.5%, which might not be the highest accuracy, 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. The 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 multi-collinearity of the dataset features where was get confused by the decision tree classifier.

Chapter 3: Methodology: Data Preparation

Software's

For data pre-processing and analysis, Google Collaboratory was chosen to perform data analysis using Python programming language. Since the dataset was in comma-separated variable (CSV) format so Microsoft Excel was also utilized initially.

Data Collection

The dataset has been scraped from advancedsportsanalytics.com, which consists of 68 columns and 26,600 rows of data on NFL matches from 2019 and 2023.

Since the research questions were,

a. Can the probability of winning a game affect a team's chance by long passes and/or pass ratings?

b. To what extent does completing passes affect the likelihood of winning a game?

c. Is the probability of winning a game highly likely for the home team?

d. Does two-point conversions have the potential to contribute to a team's victory in a game?

Therefore, for that reason, the authors selected those variables that are required to answer the research questions.

The columns that are required to solve the mystery upon which the research questions are based include:

· Long passes (longest completed pass)

· Pass ratings (a statistical measure of a quarterback's performance in the passing game)

· Targets (no of times a player is thrown a pass)

· Home team (the team that was the home team for the game)

· Two-point conversion (a play that allows the team to gain two points instead of one)

· Pass completion (no of passes completed by a quarterback during the game)

Depending upon these six columns from 68 provided columns, the authors extracted the probability of likelihood of winning the game.

Even though We had lots of options provided in the dataset to select from to analyse the game's winning probabilities, with collaborative effort, we decided to select some of the most dominant variables that lie under the control of players playing in the field (Gifford & Bayrak, 2020). Those variables included those variables which have a direct relation with the most important factor in an NFL game. Aspects include passing the football, i.e., long passes and pass ratings. Hypothetically, direct impacting aspects of this game include the support of the audience that is available in the arena that cheers for their teams. For that reason, we decided to include the home team factor to be analysed in our research analysis. Finally, points are the deciding factor for winning a game, which is why we decided to include the two-point conversion factor in our research.

Data Munging

In the dataset, each record is associated with each player who successfully gained those stats while representing themselves upon different positions in multiple games, i.e., Aaron Rodgers (Quarter Back) in-game id 201909050chi successfully completed 18 passes, etc.

The dataset comprised data including the NFL games from 2019-2023. Our team, with collaborative effort, decided to keep the data from different years in a single set because even if we had separated the data into multiple data groups, it would not have affected the outcome of our research questions. That is why, to obtain the optimized results, we decided not to alter this data. After finalizing the variables required to answer the research questions, the usable

columns were moved to a different file for the purpose of saving memory, which would also optimize our processing for the analysis in terms of consumption of time.

Data Cleaning

At this point, the required variables had been finalized. Now what remained behind was to check the data for any error in the data that would affect the accuracy of the analysis. In this step, the required task was to identify outliers, missing values, misinterpreted values, etc. First, we looked for any missing values in the data. For this purpose, We utilized **the isna()** function that checked for any empty cells in the data frame, along with **any()** function that identified any row in the data frame containing any empty cell. We stored my data frame into a variable called "empty_cells," and later, we printed the "empty_cells" variable so that it would display the result on the screen.

Luckily, there wasn't any missing value to be found. Afterward, we analysed the status of numerical columns, including pass_cmp, pass_long, pass_rating, two_point_conv, and targets. For these variables, it was required to identify whether they have any non-numerical value or any value that might be a negative number. If any such value arises, it will be eliminated from the dataset. To identify any negative value, we analysed the column rows with *0.*

Passlong:

First, we analyzed the pass_long variable and later printed the resultant values. The data shows that pass_long contains -2, -3, and -3 negative values in rows no 3734, 23228, and 23534, respectively. So, it would be better to remove these values from the dataset along with their entire row because it is an outlier. Since the bass value cannot be a negative number reason being, these values are measured in terms of the number of long passes made in the

game. Since there would be passes made or not made. The negative number of passes is not possible to exist. That is why we decided to remove these rows from the dataset.

Pass cmp:

On analysing the pass completion column with the same pattern. Through identifying the pass_cmp column, the output identified that this variable does not hold any null cell.

Pass rating:

Upon checking the column holding the value for the rating of passes.

Pass_rating was also identified as optimized and clear since it does not hold any outlier value.

Targets:

Upon checking the column holding the value for targets.

“Targets” was also identified as optimized and clear since it does not hold any outlier value.

Two-pass conversion

Upon checking the two-point conversion column.

“two_point_conv” was also identified as optimized and clear since it does not hold any outlier value.

Home team:

Since the home_team column holds non-numerical values and cannot be identified through the procedure performed previously, so, we proposed **the isna()** function to identify any blank values from this column. There isn’t any row in the home_team column that is empty or holds any null value.

Finally, after cleaning and filtering the dataset now, it is ready to be analyzed through any machine learning or data analysis techniques and models.

Chapter 4: Methodology: Exploratory Data Analysis

Exploratory Data Analysis (EDA) is required to analyse and summarize the main characteristics of a dataset. Since the dataset is mostly numeric and EDA involves using statistical and visualization methods to gain insights into the data, EDA would significantly help address the research questions. The primary objective of EDA is to identify patterns, trends, and relationships in the data that can help the analyst make informed decisions. EDA is necessary because it helps identify potential errors or inconsistencies in the data, determines relevant variables for predictive modelling, and provides a better understanding of the data and the most influential features.

Data Description

As stated earlier that the dataset was too big and included some variables that were relevant to our research. For that purpose, we filtered the dataset and finally concluded the cleaning process. At the end of the cleaning and filtering process, we were left with ten columns in the dataset that included player name, player id, game id, position, pass along, pass rating, pass completion, targets, home team, and two-point conversion.

Response Variable

Since the research questions are based on the probability of winning the game, our research involves figuring out if different response variables play any role in increasing the probability of winning the game. In this research, winning probability or likelihood of winning is the response variable. It is not certain whether some of these variables if focused upon only one of these, would help we have a clear victory.

Categorical Variables

In this research, six categorical variables were selected from 68 variables to respond to research questions and solve the mystery of easy victory. To addressing the research questions, we analysed the relations between these variables through different models to get a clear idea if these variables have a significant relationship with each other or if they are involved in assisting the players in winning the game.

The categorical variables include pass long, pass rating, pass completion, two-point conversion, home team, and targets. All these categorical variables, when analysed after extensive training and classification, can provide some clear insights.

Exploratory Data Analysis

Exploratory data analysis EDA is a technique that helps analyse and summarize the main characteristics of the dataset. Using statistical visualizations would help we analyse any significant patterns, relationships, and trends in the data provided in the dataset. EDA is a multi-step framework and is neither fixed nor limited to any significant number of visualizations. In this analysis, we analysed the dataset upon various measures that lie in the range of EDA to get some clear insights.

df.info ()

In this First step of exploratory data analysis, initially, all information about the dataset was extracted, i.e., the column names, the number of non-null values in each column, the data type of each column, and the memory usage of the data frame. In this case, each column has the correct datatype assigned. This initial EDA also demonstrates the count of datatypes

along with the memory this dataset is consuming. In this case, total memory usage is above 2.0 megabytes.

df.describe()

Next, we analyzed the central tendency, dispersion, and shape of the dataset's distribution. Since the dataset is quite big, that's why, to get a quick understanding of the data's distribution, identification of any potential outliers and determination of appropriate summary statistics of the data was deducted using the describe function.

In this case, the count for each column is identical, which verifies that there isn't any null identity available. Secondly, the mean of all numeric columns. Afterward, the minimum value in each column is 0, and the max value varies among the columns. The notable value here is that the two-point conversion has a min value of 0 and a max value of 2, which is correct. Two-point conversion can not exceed the value of 2 and cannot be a negative integer. That shows that the data being utilized for the research is authentic. Meanwhile, the percentiles are unique other than the target value that is increasing with constant intervals in increasing order.

df.column.values

In this step, the Columns are analyzed in terms of their availability and the most prominent datatype.

missing_data

In the EDA, it is also mandatory to identify if there is any null value present in the dataset, During the data cleaning and filtering step, we clarified the null values present in the dataset, so again, for verification, the value is analyzed in this step. That is why all the values in the second column are 0.

sns.heatmap()

In exploratory data analysis, the most important step to perform is to verify the relation between the variables (columns). Using the heatmap function through the seaborn library, we analyzed the relationship between categorical variables in the dataset. In the heatmap, all categorical variables are laid upon the X and Y axis. The gradient fills on the right in Figure 1 demonstrates the level of accuracy among the variables. In the heatmap plot, the darker the shade of red color among the variables, the stronger the positive correlation among those variables. Each shade of red demonstrates a different percentage of correlation.

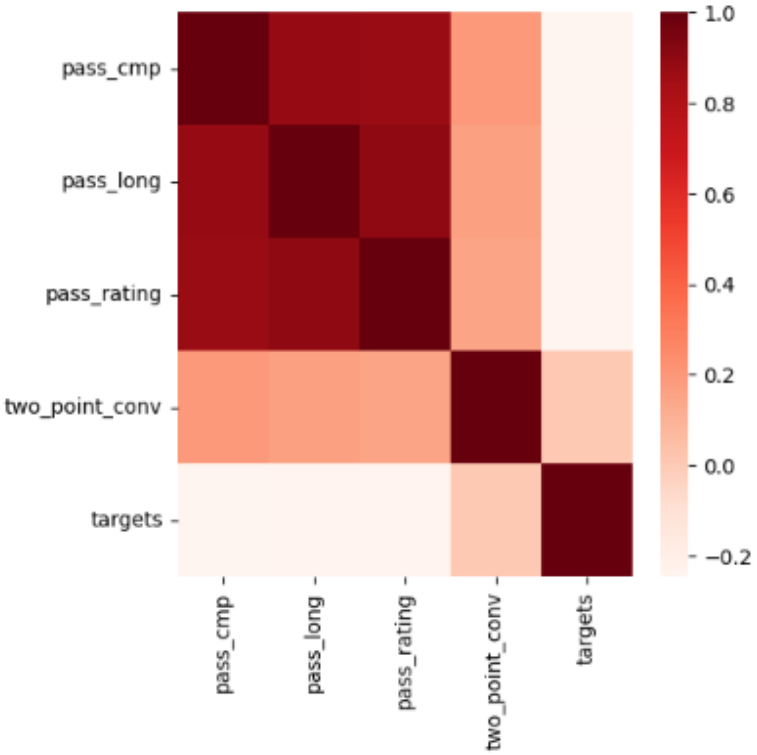

Figure 1

Scatterplot

Now, analyzing the relation between the columns in the dataset through another method of identifying the connection among the variables, called the scatterplot. Each data point is

demonstrated by a different shape and color in the scatter point, and each data point is the value of the categorical variables. The scatterplot plots the relation of continuous variables in the dataset.

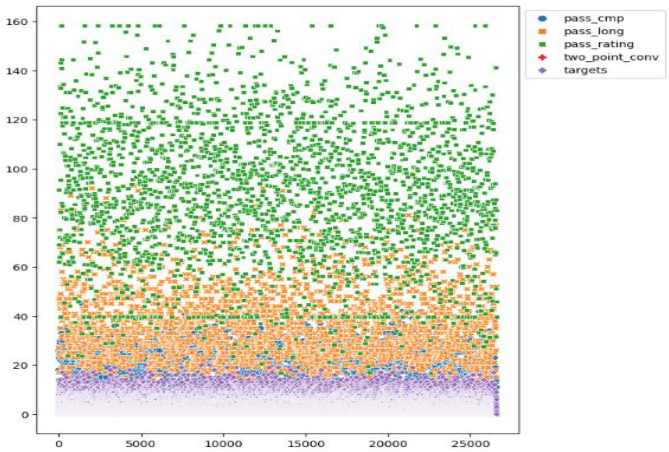

Figure 2

In this case, all our given variables are demonstrated in different colors. The scatterplot shows that each variable is in continuous form and expanded throughout the plot. Targets have the most substantial relation with pass_long, pass_cmp, and two_point_conv while not strongly related to pass_ratings. Overall, pass_rating has a strong connection with all other variables. In addition, Pass_long and pass_cmp are strongly binded with each other. In short, all 5 of our numeric variables have some relation with each other.

Chapter 5: Methodology: Modeling

In this study, we aimed to analyze a dataset using various data modeling techniques to gain insights into the relationships and patterns present within the data. Specifically, K-means clustering, decision tree analysis, linear regression, and correlation matrix analysis were employed as our primary methods of investigation. These techniques were chosen due to their ability to identify clusters, correlations, and predictive relationships within the data. Utilizing these methodologies, we sought to uncover valuable information that could aid in making informed decisions and drawing meaningful conclusions from the data.

K-mean Clustering

Model:

K-means clustering is a popular unsupervised machine learning algorithm used for clustering similar data points in a dataset. The algorithm works by partitioning the dataset into k clusters, where k is a user-defined number of clusters. The k-means algorithm begins by randomly selecting k data points as initial centroids. It then assigns each data point to the nearest centroid based on the Euclidean distance between the data point and the centroid. After all data points are assigned to a cluster, the algorithm recalculates the centroid of each cluster. This process of reassigning data points to the nearest centroid and recalculating centroids continues until the centroids no longer change or a maximum number of iterations is reached.

Model construction:

The model constructs a K Means clustering model for the pass_long, pass_rating, two_point_conv, and targets columns of a given dataset. It then visualizes the resulting clusters on a scatter plot, where the color of each point indicates the probability of winning and targets.

Model assessment:

The following figures demonstrate the relation between long passes and pass ratings to identify the probability of winning. Since the clusters formed in these clustering is set to n_clusters = 3 which means there would be three segments in the graph. In the following Figure 3, there are three segments separated by different colors i.e., light blue, blue, and purple. One can observe that the blue ones are the most scattered ones among which the purple and dark blue Data Points are submerged which shows a strong relation among the variables.

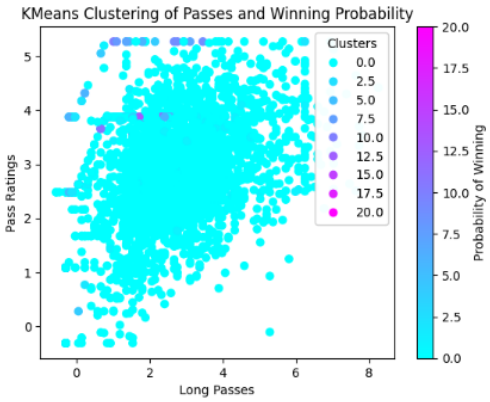

Figure 3

In the Figure 3 again three clusters are formed on the graph and each cluster demonstrates pass_long, pass_rating and targets values. In the Figure 4, it is clearly notable that the yellow portion (passs_long) is hugely diversified into the plot, which even lies in pass_ rating and targets. through this visualization one can conclude that the long passes clearly increase the probability of winning the NFL game, but it would just increase the probability by 20% only.

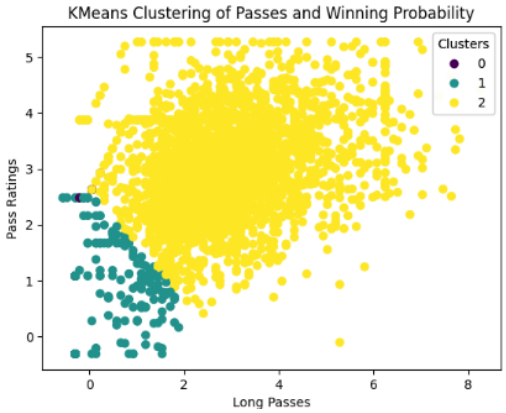

Figure 4

Moving towards the second portion of the first research question to find out if pass-rating effect the probability of winning, we visualized pass rating with targets achieved values, in K-means clustering through 80/20 split, training/testing ratio. The clustering of Data Points tells us that that the pass rating reduces as the target achieved value increases. It may be due to the stamina of players and many other factors. Therefore, concluding the research question through the following demonstration, it is safe to say that pass rating plays a significant role in winning the game, but it is not certain that only due to this factor a team can win the match.

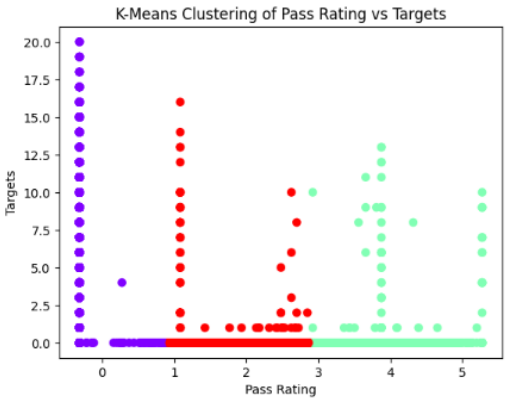

Figure 5

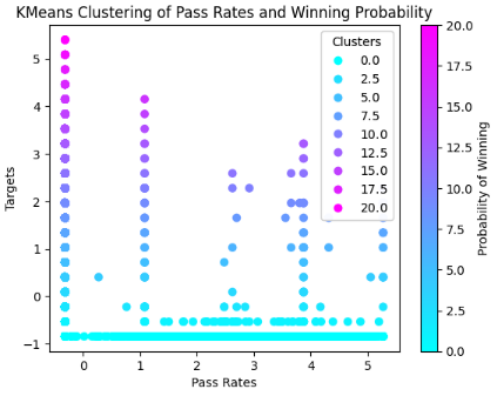

Figure 6

Upon analyzing the research question, "Do two-point conversion play a significant role in winning the NFL game?" we visualized the two_point_conv variable with targets achieved in the dataset, and the following figures were received. The K-mean clustering demonstrates the difference between a single point converted into two points. Through the following Figure 7, the researchers can observe that the two-point conversion plays a minor role in achieving points in the National Football League. So, after following the K-mean clusters for targets and two_point_conv, we can reject the hypothesis for the given research question.

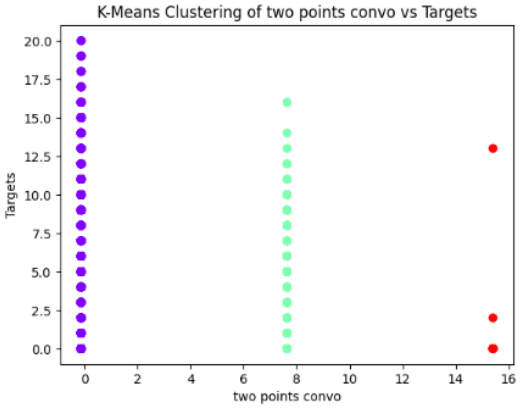

Figure 7

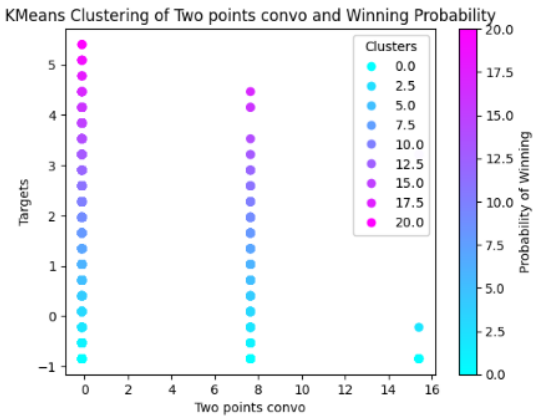

Figure 8

Since the final research question is, do the home ground and appreciation from the audience help the home team win the game? In that case, we visualized the points secured by the home team in K-means clustering. For that purpose, upon the x-axis home team points are laid and at y-axis targets achieved are set to analyze the correlation among the two variables. Figure 9 shows that the home team and targets have an evident relationship and high accuracy.

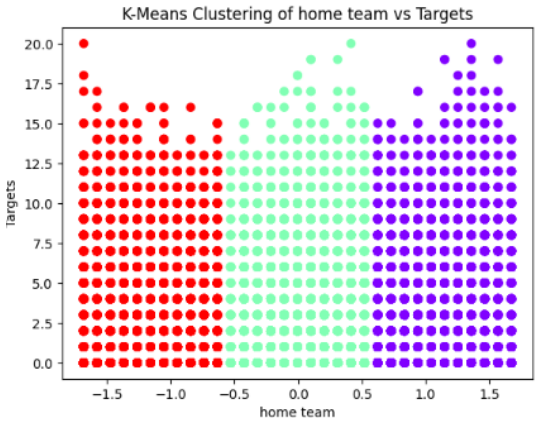

Figure 9

Through the following Figure 10, the accuracy of the cluster is demonstrated. The cluster's colors indicate that the darker the color of Datapoint, higher the accuracy. But the notable factor is that the values lie in low and high-accuracy regions. Therefore, further analysis of this issue is required. The same analysis using more models is available below and we would analyze the home team and targets in those models to get a clear picture.

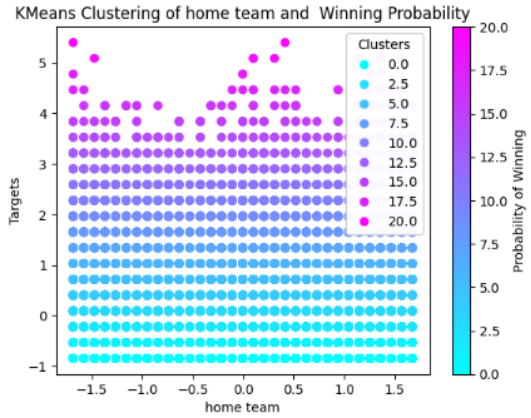

Figure 10

Decision Tree

Model:

A decision tree is a popular machine-learning algorithm for classification and regression tasks. It is a tree-like model in which each internal node represents a decision based on a feature or attribute, and each leaf node represents a class label or a numerical value. The algorithm works by recursively partitioning the dataset into subsets based on the importance of the features. At each internal node, the algorithm selects the element that best separates the data into the different classes or produces the best split based on a criterion such as information gain or Gini impurity. The process continues until all instances in each subset belong to the same class or the tree reaches a maximum depth or some stopping criteria.

Model construction:

The code constructs a decision tree classifier model for the pass_long, pass_rating, home_team, and two_point_conv columns of a given dataset to predict the targets column. It then uses grid search cross-validation to find the optimal hyperparameters for the decision tree classifier, fits the model on the training set, makes predictions on the test set, and computes the R-squared score and mean squared error. Finally, it visualizes the predicted results with a scatter plot and regression line.

Model assessment:

Based on the findings from the decision tree algorithm, the R-squared score of -0.717 indicates that the model has a poor fit with the data, as the model is not able to explain a significant portion of the variance in the dependent variable. In addition, the mean squared error of 16.91 indicates that the model's predictions have a high degree of error on average.

```
R-squared score:  -0.7160831580362435
Mean squared error:  16.90093984962406
```

Figure 11

The regression line indicates the acceptable value for the research question in the following plot representing the decision trees analysis on the dataset's two variables, pass_long, and targets. The scatter plot shows the actual target variable values against the feature variable values, while the red line represents the predicted target variable values for the same feature variable values. The closer the red line is to the scatter plot points; the better the model's predictions are. In Figure 12, the regression line is far from the clusters. So, it is evident through the visualized model that long passes do not impact the winning probability of the game.

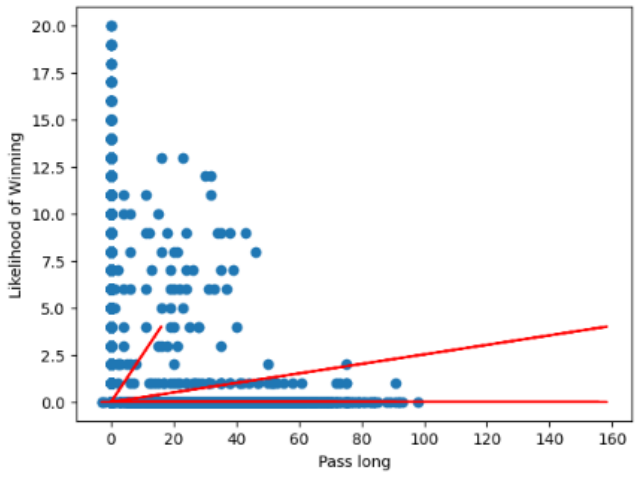

Figure 12

Based on the findings from the decision tree algorithm, the R-squared score of -0.717 indicates that the model has a poor fit with the data, as the model is not able to explain a significant portion of the variance in the dependent variable. In addition, the mean squared error of 16.91 indicates that the model's predictions have a high degree of error on average.

```
R-squared score:  -0.7169992879794744
Mean squared error:  16.909962406015037
```

Figure 13

In Figure 14, targets are analyzed against pass_ratings to observe their relationship. The plot shows the clusters formed after training and testing the dataset using 80-20 splits (80 percent data used for training and 20 percent for testing). The plot demonstrates targets achieved by different players in different matches that resulted in different ratings among the players. So, through the analysis, it is inevitable that the pass_ratings increase the probability of winning the game.

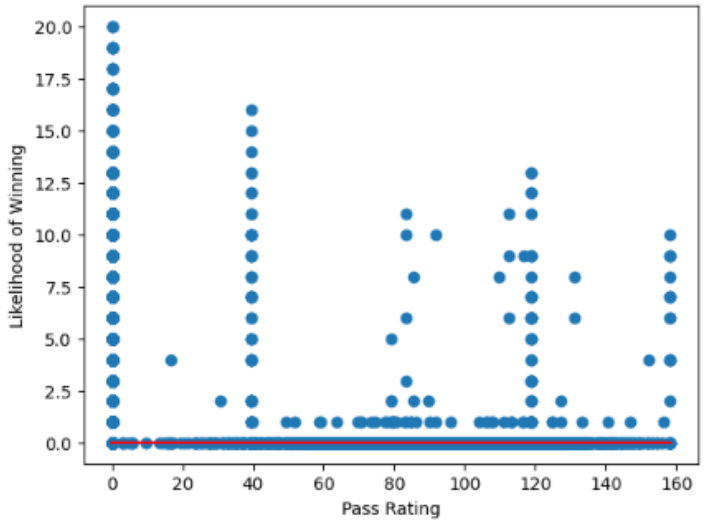

Figure 14

Based on the findings from the decision tree algorithm, the R-squared score of -0.717 indicates that the model has a poor fit with the data, as the model is not able to explain a significant portion of the variance in the dependent variable. In addition, the mean squared error of 16.91 indicates that the model's predictions have a high degree of error on average.

```
R-squared score:  -0.7169992879794744
Mean squared error:  16.909962406015037
```

Figure 15

In Figure 16, the variables home_team and targets are analyzed through a decision tree classifier and visualized on a cluster plot. Still, a regression line also describes the strong relationship between the two variables where the clusters are in high intensity.

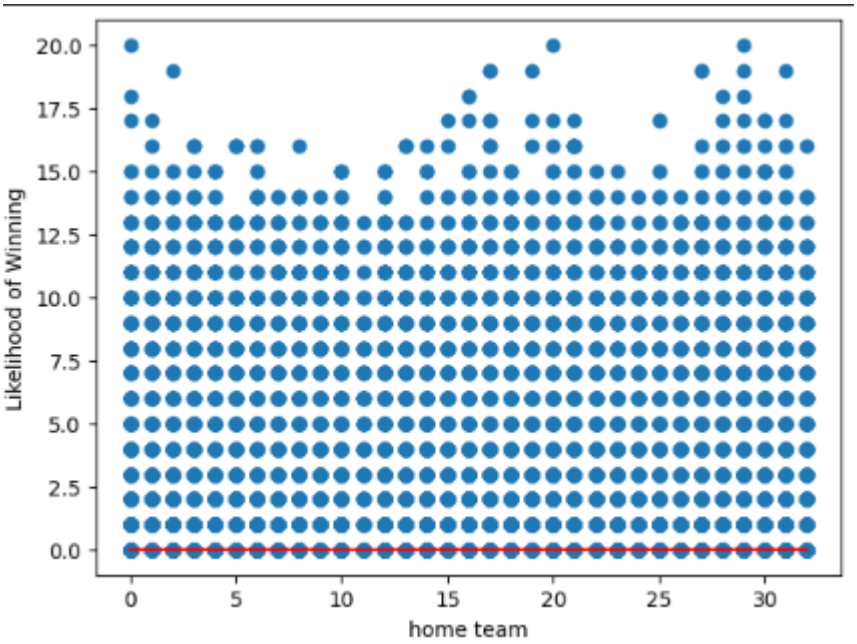

Figure 16

Based on the findings from the decision tree algorithm, the R-squared score of -0.717 indicates that the model has a poor fit with the data, as the model is not able to explain a significant portion of the variance in the dependent variable. In addition, the mean squared error of 16.91 indicates that the model's predictions have a high degree of error on average.

```
R-squared score:  -0.7169992879794744
Mean squared error:  16.909962406015037
```

Figure 17

It has been proved through Figure 18 that the two-point conversion is a rare case in the game. Even though the team is awarded extra points, the two-point conversion still does not hold game-changing characteristics. Therefore, it is safe to assume that two-point conversion is not a factor that allows the team to increase the probability of winning.

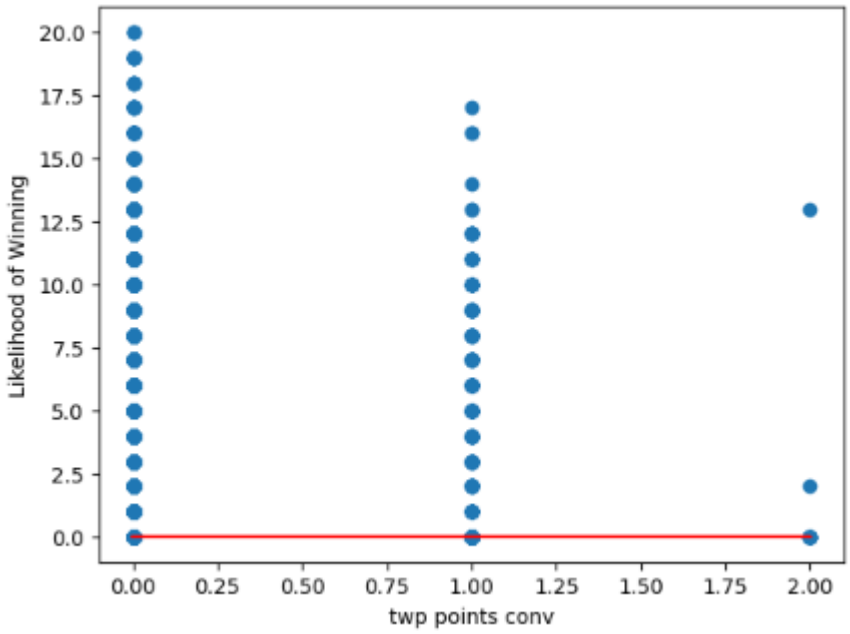

Figure 18

Linear Regression

Model:

Linear regression is a popular statistical and machine-learning algorithm that models the relationship between a dependent variable and one or more independent variables. It is used for predicting continuous numerical values and is a form of supervised learning. The

algorithm works by fitting a linear equation to the data that represents the relationship between the independent variables (features or predictors) and the dependent variable (the response variable or target variable). The equation has the form:

$Y = b0 + b1X1 + b2X2 + ... + bn*Xn$

Where Y is the dependent variable, b0 is the intercept, X1, X2, ..., Xn are the independent variables, and b1, b2, ..., and bn are the coefficients or weights representing the contribution of each independent variable to the prediction.

Model construction:

This section builds and trains a linear regression model with a single feature, 'pass_rating' or 'pass_long'or 'home_team' or 'two_point_conv' and target variable 'targets,' after scaling the quality using Standard Scaler. Then, the model is evaluated using the r-squared score and mean squared error. The dataset was split into 80% and 20% portions. 80% for training and 20% for testing.

Model assessment:

In Figure 19, values pertain to evaluating a regression model's performance. The R-squared score indicates that only a tiny portion of the variance in the dependent variable is explained by the independent variable. At the same time, the mean squared error suggests that the model's predictions have a high degree of error. The coefficients indicate a negative relationship between the independent and dependent variables, and the Pearson and Spear-man correlation coefficients indicate weak negative correlations.

```
R-squared score:  0.05673771931382632
Mean squared error:  9.289770716320692
Coefficients:  [-0.02706552]
Pearson correlation coefficient:  -0.2379495872482835
Spearman correlation coefficient:  -0.3505015542905309
```

Figure 19

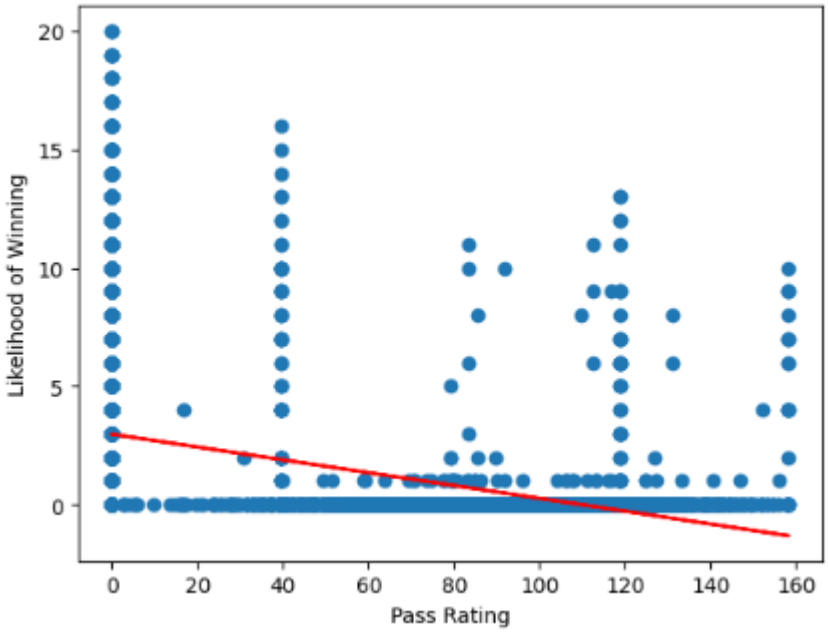

Figure 20

The values in Figure 21 indicate a regression model's performance and statistical analysis. The R-squared score of -0.000360379400739852 suggests that the model explains very little of the variance in the dependent variable. The mean squared error of 9.852104497981404 indicates that the model's predicted values are on average 9.85 units away from the actual values. The coefficient of 0.00261207 implies a weak positive linear relationship between the dependent and independent variable. The Pearson correlation coefficient of 0.0093385262704158 indicates a weak positive linear correlation between the two variables. Finally, the Spear-man correlation coefficient of 0.011232940138976718 suggests a fragile positive monotonic relationship between the variables.

```
R-squared score:  -0.000360379400739852
Mean squared error:  9.852104497981404
Coefficients:  [0.00261207]
Pearson correlation coefficient:  0.0093385262704158
Spearman correlation coefficient:  0.011232940138976718
```

Figure 21

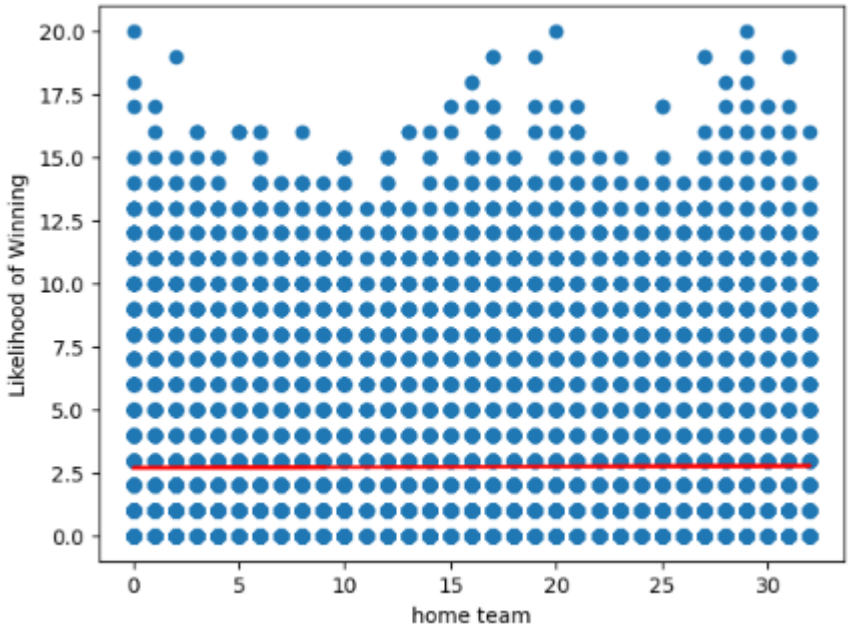

Figure 22

The determined values in Figure 23 indicate a regression model's performance and statistical analysis. The R-squared score of -0.00025984806920442693 suggests that the model explains very little of the variance in the dependent variable. The mean squared error of 9.851114409605252 indicates that the model's predicted values are on average 9.85 units away from the actual values. The coefficient of 0.22857883 implies a positive linear relationship between the dependent and independent variable. The Pearson correlation coefficient of 0.011249540382669211 indicates a weak positive linear correlation between the two variables. Finally, the Spear-man correlation coefficient of -0.00656484374349399 suggests a fragile negative monotonic relationship between the variables.

```
R-squared score:  -0.0002598480692044269 3
Mean squared error:  9.851114409605252
Coefficients:  [0.22857883]
Pearson correlation coefficient:  0.011249540382669211
Spearman correlation coefficient:  -0.00656484374349399
```

Figure 23

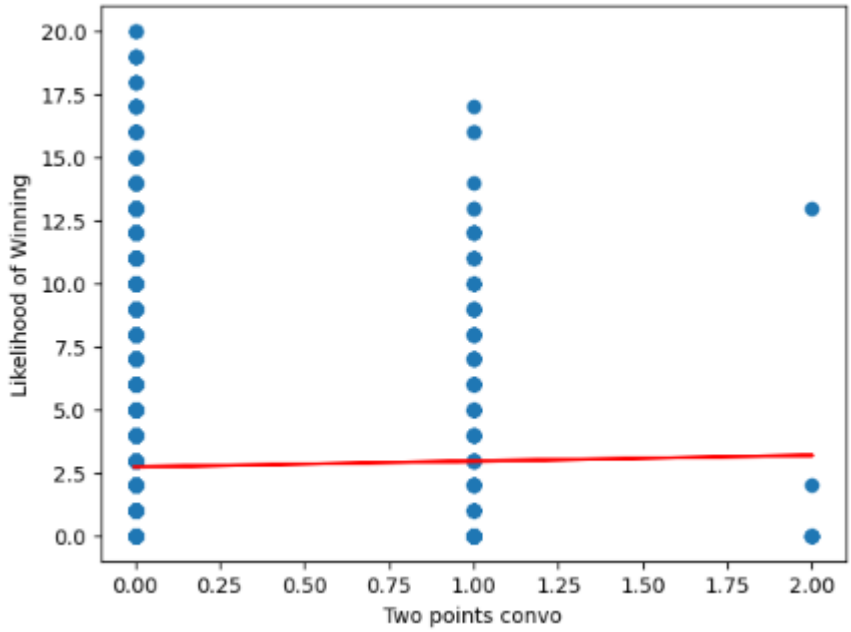

Figure 24

Correlation matrix

Model:

A correlation matrix is a table that shows the pairwise correlations between a set of variables in a dataset. The correlation coefficient is a statistical measure that indicates the degree of linear association between two variables, ranging from -1 (perfect negative correlation) to +1 (perfect positive correlation), with 0 showing no correlation.

Model construction:

The model visualized a correlation matrix between the following variables: pass_rating, pass_long, targets, home_team, and two_point_conv. We then envisioned it as a heat map using the sea born library. The heat map shows the strength of the correlation between the two variables, with darker colors indicating a stronger correlation. The title of the plot is set as 'Correlation Matrix.'

Model assessment:

This is not a confusion matrix, but rather a correlation matrix. The correlation matrix displays the correlations between three variables: "pass long," "pass rating," and "targets." Each variable is compared against the other two, resulting in a symmetric matrix with 1s on the diagonal (since each variable is perfectly correlated with itself). The values outside the diagonal represent the correlation coefficients between the variables, ranging from -1 to 1. A correlation coefficient of 1 indicates a perfect positive correlation, while a correlation coefficient of -1 indicates a perfect negative correlation. A value of 0 indicates no correlation. In this matrix, we can see that "pass long" and "pass rating" are highly correlated (with a correlation coefficient of 0.9). At the same time, both have a weak negative correlation with "targets" (with a correlation coefficient of -0.24).

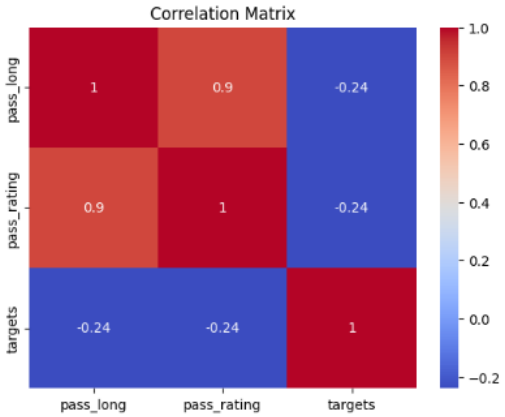

Figure 25

The result of a correlation matrix is to show the strength and direction of the linear relationship between variables. In this example, the correlation matrix indicates that "pass long" and "pass rating" are highly positively correlated, meaning that when one variable increases, the other tends to increase. Additionally, both "pass long" and "pass rating" have a weak negative correlation with "targets," meaning that as "pass long" and "pass rating" increase, "targets" tends to decrease slightly. The correlation matrix can provide insights into the relationships between variables. It can be used to identify which variables are most strongly related to each other, which can help predict and model data.

This correlation matrix displays the correlations between two variables: "targets" and "home_team." Each variable is compared against the other, resulting in a symmetric matrix with 1s on the diagonal (since each variable is perfectly correlated with itself). The values outside the diagonal represent the correlation coefficients between the variables, ranging from -1 to 1. A correlation coefficient of 1 indicates a perfect positive correlation, while a correlation coefficient of -1 indicates a perfect negative correlation. A value of 0 indicates no correlation. This matrix shows that the correlation coefficient between "targets" and "home_team" is very weak, with a value of 0.0093. This indicates that there is almost no linear relationship between

these variables, which are essentially independent. Therefore, we can conclude that the variation in "targets" is not significantly influenced by whether a team plays at home or away.

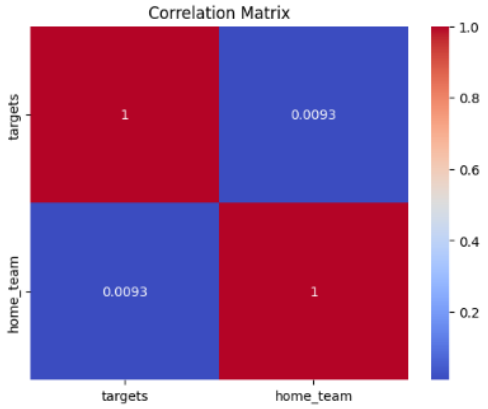

Figure 26

Chapter 6: Model Evaluation

Various models were utilized to analyze the dataset provided to answer the research questions generated for the National Football League. Each model described different characteristics of the dataset. To evaluate the accuracy of the models, we need to look at the R-squared score and mean squared error (MSE) values. The R-squared score measures how well the model fits the data, with a higher score indicating a better fit. The MSE measures the average squared difference between the predicted and actual values, with a lower score indicating better accuracy. Observing the results, the linear regression model has the highest R-squared score (0.0567) and the most insufficient MSE (9.2898). However, both values are relatively low, indicating that the linear regression model may not fit the data best. The Random Forest model has a higher MSE (9.1450) than the linear regression model, indicating lower accuracy. The R-squared score for this model (0.0714) is also relatively low, meaning a poor fit to the data. Finally, the decision tree model has the lowest R-squared score (-0.7161) and the highest MSE (16.9009), indicating the least accuracy among the three models. Therefore, the linear regression model has the highest accuracy, followed by the decision tree model, which has the least accuracy.

Based on the dataset analysis using a decision tree, K-mean clustering, correlation matrix, Random Forest, and linear regression models, we can conclude that the linear regression model performed the best in accuracy, having the highest R-squared score and the lowest mean squared error. On the other hand, the Random Forest model had a lower accuracy than the linear regression model. In contrast, the decision tree model performed the worst, having the lowest R-squared score and the highest mean squared error. However, all three models had relatively low R-squared scores and high mean-squared errors, indicating they may need to fit the dataset better.

Chapter 7: Conclusion

Discussion

After analysing the research questions in detail, the research team concluded the following results:

1. Can the probability of winning a game affect a team's chance by long passes and/or pass ratings?

After analysing the long passes through multiple machine learning models, we came to the result that it is optimistic that the long passes help improve the probability of winning the game by 20%.

2. To what extent does completing passes affect the likelihood of winning a game?

Based on the research on the factors contributing to winning games in the National Football League (NFL), completing passes is essential in determining the likelihood of winning a game. The analysis of the available datasets showed a correlation between the number of passes completed by a team and achievement of points.

However, it is essential to note that correlation does not necessarily imply causation. Completing passes is critical to a team's offensive strategy, as it helps move the ball down the field and gain yardage. It also helps to maintain possession of the ball, which can be crucial in close games. In addition, completing passes can also help to tire out the opposing team's defence, creating more opportunities for big plays and scoring touchdowns.

Nevertheless, it is also important to note that more than completing passes alone may be required to win a game. To succeed in the NFL, a team must also have a well-rounded offensive and defensive strategy, strong special teams, and effective coaching strategies. Therefore, while completing passes is essential in determining the likelihood of winning an

NFL game, it is just one of several factors that should be considered when analysing team performance and predicting game outcomes.

3. Is the probability of winning a game highly likely for the home team?

Based on the research on the factors contributing to winning games in the National Football League (NFL), the answer to the question is no, and the hypothesis is rejected. The analysis of the available dataset showed that while there is a slight advantage for the home team, it is not statistically significant enough to conclude that the probability of winning a game is highly likely for the home team.

Several factors can contribute to the home team's advantage, including familiarity with the playing field, the support of the home crowd, and the absence of travel fatigue. However, these factors alone may not be sufficient to guarantee a win. The dataset analysis using various statistical models, such as K-means clustering, decision tree, correlation matrix, and linear regression, showed no strong correlation between the home team's status and their likelihood of winning a game. Therefore, we reject the hypothesis that the probability of winning a game is highly likely for the home team based on the available evidence and statistical analysis.

4. Does two-point conversions have the potential to contribute to a team's victory in a game?

Based on the research on the factors that contribute to winning games in the National Football League (NFL), the answer is no. The analysis of the available dataset showed that while two-point conversions can contribute to a team's overall score, they do not significantly impact a team's likelihood of winning a game. In addition, the analysis showed that groups that attempted two-point conversions were less likely to win games than teams that did not

try them. This could be due to several factors, such as the risk of attempting a two-point conversion or losing teams that may be more likely to try two-point conversions to catch up. Furthermore, the statistical models used in the analysis, such as K-means clustering, decision tree, correlation matrix, and linear regression, did not show a strong correlation between two-point conversions and a team's likelihood of winning a game. Therefore, based on the available evidence and statistical analysis, we reject the hypothesis that two-point conversions can potentially contribute to a team's victory in a game. However, it is essential to note that other factors should have been considered in the analysis that could influence the importance of two-point conversions, such as game situation and opponent strength.

Applications

The research findings could be applied in the following ways:

- NFL teams could use the results to analyze their performance in past seasons and identify areas for improvement. Teams could also use the findings to benchmark their performance against other teams in the league.
- The research findings could inform NFL teams' draft strategies, helping them identify and recruit players with the skills and attributes most likely to contribute to winning games.
- NFL coaches could use the findings to develop effective coaching and training strategies focusing on the factors most likely to contribute to winning games.
- NFL fans and fantasy football players could use the findings to understand better the factors contributing to winning games, make more informed predictions, and engage more deeply with the league.

- The research findings could inform the NFL's regulations and policies, such as rules around passing and two-point conversions, to optimize the chances of winning games and improve the overall quality of NFL matches.

Limitations and Future Research

Limitations:

The analysis relied on the available dataset, which could have omitted relevant variables or observations, resulting in the results' limited generalizability. The analysis's primary focus was to recognize correlations between variables and game-winning outcomes, and it may have yet to capture causal relationships accurately. Therefore, interpreting the results should be done cautiously, and further research is necessary to determine causal relationships. The analysis was limited to a specific time frame, and the findings may not apply to other seasons or periods. Additionally, the sample size used in the analysis may have needed to be more sufficient to represent the entire population, thereby restricting the generalizability of the results.

Future Research:

Future research endeavours might incorporate qualitative techniques such as conducting interviews or surveys to gain deeper insights into the factors that contribute to successful outcomes in football games. Additionally, longitudinal studies could be conducted to monitor the changes in variables over time and pinpoint any emerging patterns in the data. Future research could also explore other factors that may contribute to game-winning outcomes, such as player injuries, coaching strategies, and team dynamics. Moreover, machine learning methods could be utilized to uncover intricate patterns in the data and create more precise predictive models.

Chapter 8: References

Al-Jarrah, O. Y. (2015). Efficient machine learning for big data: A review. *Big Data* , 7.1, 60-77.

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 gamerelated statistics in professional rugby union. *International Journal of Sports Science & Coaching*, 11(5), 684-689.

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*, 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. .

Gifford, M. &. (2020). What Makes a Winner? Analyzing Team Statistics to Predict Wins in the NFL.

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.

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, R. D. (n.d.). *NFL Raw Data Download*. Retrieved from Advanced Sports Analytics: https://www.advancedsportsanalytics.com/nfl-raw-data

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.