NFL: A Past, Present and Future

Members: Roshan Peri (peri.ro@northeastern.edu)

Background and Problem Statement:

With the addition of new teams, extra games and how the game itself is changing, the NFL is an ever-growing field. With these changes, what are the main differences and similarities that help separate the NFL of the past from the NFL of today? That is the main question that will be addressed throughout this project: Over the past 25 years, how has the NFL changed and in what ways has it changed, and can these changes help accurately predict the future? Can some statistical categories effectively help predict the number of wins a team can have in the season? Some of the audiences that this study can help benefit is NFL general managers, team owners, as well as NFL analysts as this can give some insight into what areas can be improved and what areas are the strengths based on the team wins and the correlations based on different categories and numbers of wins. With some more in-depth research as well as access to more data, such as the players of the team during each season and their individual statistics, this research can even help predict how the addition of a player to a team can help predict the success of that team for the future seasons.

Introduction to Data:

The data that was used in this analysis was a CSV filed called ‘nfl-team-statistics.csv’ from SCORE Sports Data Repository. The creator of the dataset collected the data using the nflreadr package in R, which pulls the data from a GitHub repository. Some of the ethical concerns that could be present with this can be attributed to the following:

mental health impact of the players, the players personal information being leaked. Since the NFL is widely criticized by the fans, analysists and the coaches themselves, this data could lead to negatively impacting the players mental health, especially if they are being the one criticized for the play or if the team they play for is being criticized based off their statistics. Another ethical concern is the idea that the player's information is publicly available, such as where the player is from and their high school. This could lead to some potentials of unwanted interactions with fans. Some bias that could be present in the data is the idea that the statistics csv file does not give all the statistics that are available such as if the team made the playoffs, how many penalties each team committed throughout the year.

Data Science Approaches:

The data science approaches that were used throughout the analysis include: Simple Linear Regression, Multiple Linear Regression, KMeans Clustering as well as PCA. Simple Linear Regression is a process that finds the line of best fit that mostly fits the data that is given so it can be used to predict values for given inputs. The way this approach works is that one first looks at the correlation among the variables in the dataset. Correlation essentially helps to determine if there is a relationship between the two variables. One metric that helps with identifying correlation is the Pearson Coefficient, a value from -1 to 1 where the closer the value to -1 and 1, the more accurate the line of best fit will be. Next, the line of best fit is computed, getting back the slope and intercept. With this line, one plots all the x-values and the y-values that are computed are the predictions. One then plots these predictions to help visualize the line of best fit. The next technique that is used is a more advanced version of Simple Linear Regression called Multiple Linear Regression. This technique is a type of supervised machine learning, which essentially is a predictive model. One of the key characteristics of a supervised machine learning technique is that the data is split into two categories: training data and test data. The training data, which is composed of around 70% of the data, is the data where one knows some of the x values and y values. The rest 30% is known as the test data where you get a new x value, and one must predict the y values. One difference between multiple and simple linear regression is that with simple linear regression, one can only predict the y values based on one independent variable, hence the formula of y=mx+b. With multiple linear regression, the formula allows for more than one independent variable with the equation being y=b1x1 + b2x2 + … + bnxn + c. The next machine learning technique that was used is KMeans clustering. KMeans clustering is a type of unsupervised learning. Unsupervised learning is a type of mechanism where there are no labels on the results so one does not truly know where exactly the data belongs. This method just helps to group the data based on the similarities that are present. The way this mechanism works is that one first has to decide on the optimal k value, with k being the optimal number of clusters. They then would choose the K centroids, which consists of randomly-chose points. Then one has to assign each point to the closest centroid. After that, one has to re-compute centroids and then repeat this clusters are stable. Another machine learning technique that was used in combination with KMeans is PCA (Principal Component Analysis). PCA is a technique where one of its main goals is to reduce the dimensionality of a dataset while preserving the most important patterns or relationships between the variables without any prior knowledge of the target variables (GeeksForGeeks). It is used to examine the interrelations among a set of variables (GeeksForGeeks).

Results and Conclusion:

One of the main observations and main questions that was answered during the analysis, and was talked about in the presentation was: how has the NFL changed in the past two decades. The other question, and the focus here, was: can some statistical categories effectively help predict the number of wins a team can have in the season? The main takeaway here is that it's very hard to predict the number of wins that a team has accurately. With the simple linear regression, I found out that the best metric to determine the number of wins is basing it off how many points the team scores. One of the analyses that I did was I took six random teams (the teams were randomly generated with ChatGPT), the Miami Dolphins, the Dallas Cowboys, The New England Patriots, The Detroit Lions, The Baltimore Ravens and the Philadelphia Eagles, and performed a simple linear regression to try and find the best independent variable to help predict the dependent variable, number of wins. With the simple linear regression, with the x value being the points scored and the y value being the number of wins, the R-value was around 0.71 with the regression plot shown in figure 1. Even though this was based on the six teams and not all the teams, we get similar numbers with all 32 teams. With all 32 teams, the r value was around 0.73 with the regression plot shown in figure 2. With other independent variables, the correlation was not as well pronounced as this. For example, with the independent variable being the total amount of offensive yards, the r value was around 0.52 for all 32 teams, and with the independent variable being the number of yards given up by the defense, the r value was around -0.32 showing a very weak negative correlation for all 32 teams. The next analysis that I did was a multiple linear regression to see if multiple of these categories can be used to predict the number of wins. For the six teams, I performed this analysis with the independent variables being: 'offense_total', 'total_plays', 'takeaways',' turnovers', 'defensive_total', 'points_scored'. With the Eagles, for example, the model was a somewhat accurate model, with Philadelphia getting a root mean square error (RMSE), which is a metric to measure the accuracy of a linear regression model, of around 2.90. The graph of the predicted number of wins vs the actual number of wins is shown in figure 3. One of the main observations here is that, although the model generally follows the trends of the wins, it overestimated the early years, which in turn threw off the rest of the predictions. The next team that is worth looking at is the Baltimore Ravens. The Ravens ended with an RMSE score of around 1.83. The graph in figure 4 shows that while there are some incorrect spikes and dips in the predictions, it's an overall accurate model. The rest of the teams RMSE are as follows: New England = 3.49, Miami = 2.42, Detroit = 1.54, Dallas = 3.57. The next analysis that was performed was a KMeans clustering, with the same variables that were used in the multiple linear regression. I proceeded to do it for the 6 NFL teams as well as all 32 teams (the 6 teams were highlighted in the presentation). The first part, however, was generating an elbow graph (figure 5) to find the optimal K-value, which was found to be around 4. With this, I proceeded to do a do a KMeans clustering (with PCA being applied to the data before clustering) and with figure 6. This graph shows that, although it's a very loose relationship, there still is a relationship between these categories and the teams showing that they can be categorized, nonetheless. After all these analyses, one key takeaway is that: it is very hard to predict the NFL and how each team will perform. With the simple linear regression, it is evident especially with the wide-ranging R-values. This then suggests that there is not one good metric to help determine what closely correlates to the number of wins that a team can except. With the multiple linear regression, it is even more evident that it is hard to predict all the teams, as was shown with the wide ranging RMSE values. This shows that even with many different statistical categories, the NFL is so unpredictable just based off small things, such as penalties, missed field goals, missed extra points and many more. A team can have a statistically amazing day, with a lot of passing yards, lot of points scored, no turnovers and a few takeaways, but still lose because of a missed extra point or a field goal or a penalty which nullified a game winning score. It might work for some of the teams, but we cannot expect the same outcome with all the teams. With the KMeans clustering, even though there is some relationship between all the teams, it's unclear what the clustering is based off due to the nature of unsupervised learning.

Future Work:

Some future implications that can be taken from this analysis is that, while it was found that accurately predicting NFL wins and season outcomes is very unpredictable due to all the external factors explained above, it can still give you a sense of where the team is heading. With this type of analysis, a team could really dive into where their weaknesses lie and look to improve that in the offseason by either acquiring new players through free agency or the draft. Something that I would do would, is find a dataset of all the draft prospects and their college statistics and place them on a team where they need that specific position and try and predict how that player would affect the team, either in a positive or negative way.

Figures (all other plots can be found in the code and some in the presentation):

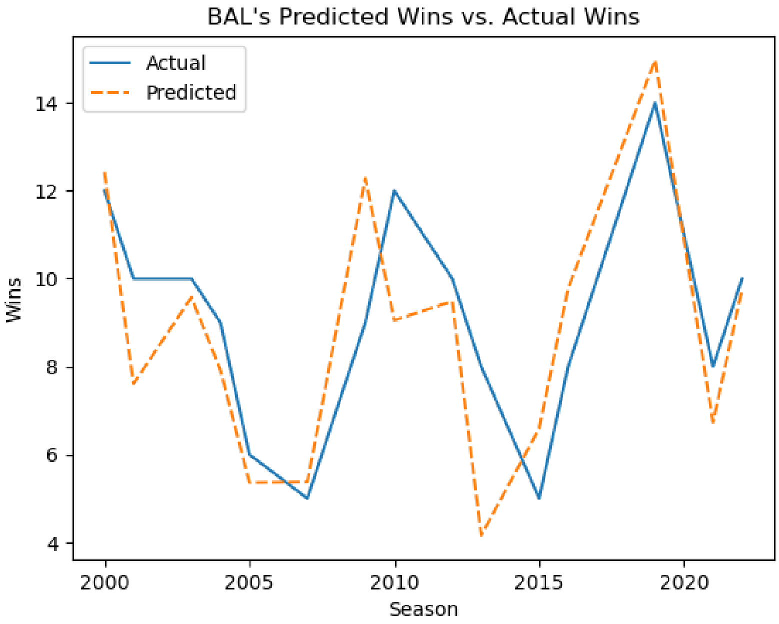

Figure 4: Baltimore's Predicted vs Actual Wins

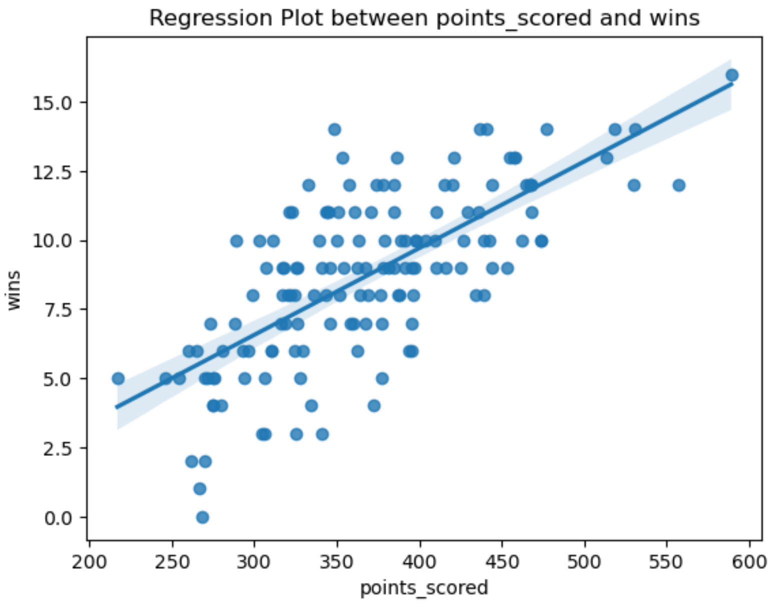

Figure 1: Regression Plot for the six teams

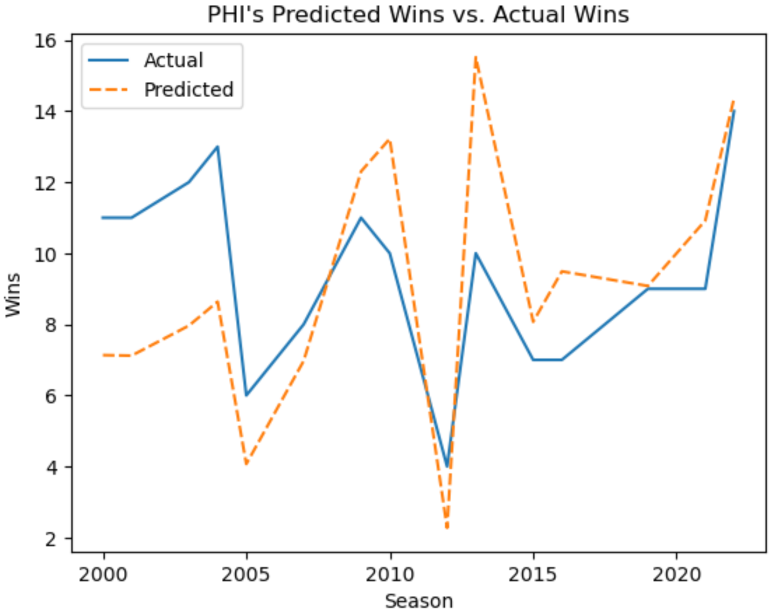

Figure 3: Philadelphia's Predicted vs Actual Wins

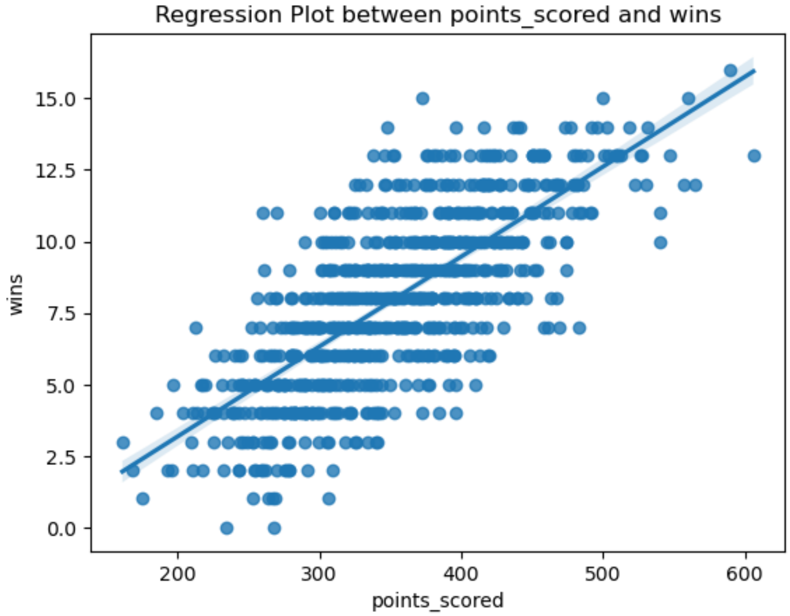

Figure 2: Regression Plot for all 32 teams

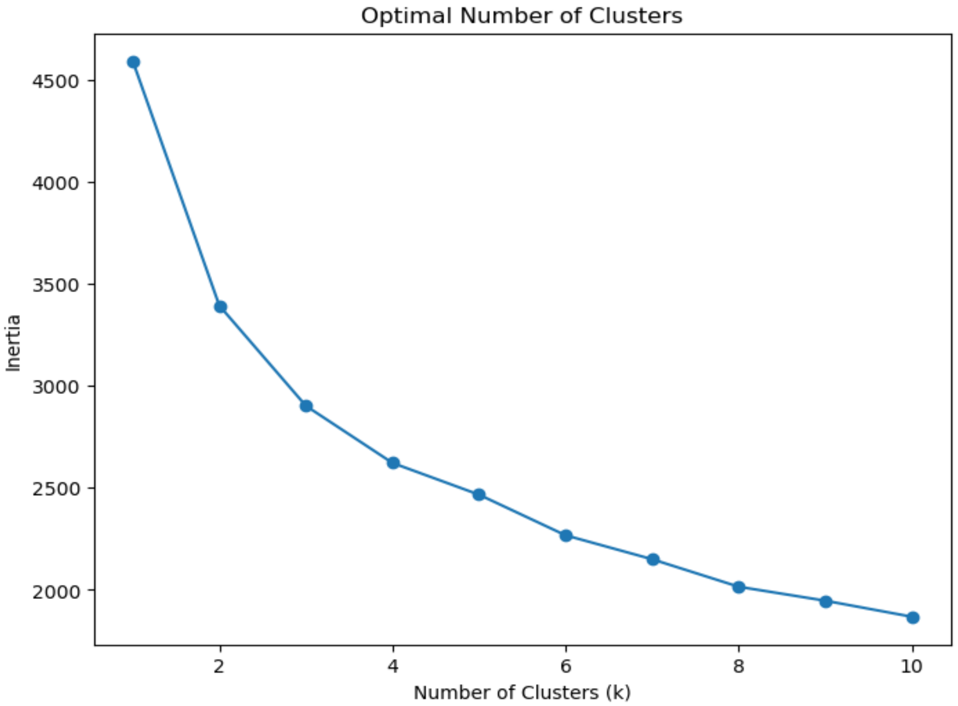

Figure 5: Elbow Plot to Find Optimal K

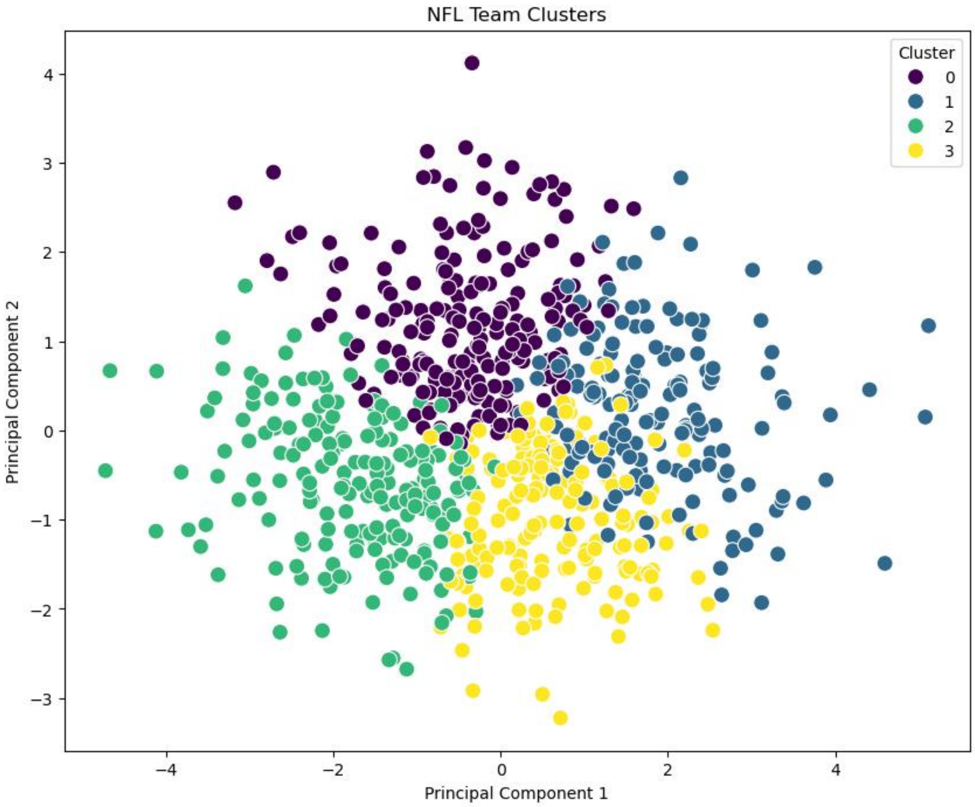

Figure 6: KMeans Clustering Plot on all 32 Teams

References:

Most of the techniques described was information taken from the in-class slides

PCA: https://www.geeksforgeeks.org/principal-component-analysis-pca/

Dataset: https://data.scorenetwork.org/football/nfl-team-statistics.html