## Machine Learning Analysis

When it came to the machine learning aspect of this project, my initial focus was on a Logistic Regression model. Although many of these statistics deal with continuous variables, the results are essentially binary; teams either win or lose (I did not really consider teams tying because of the rarity of its occurrence). And since the clubs were split between Road and Home teams, I essentially did the same model twice with each focusing on one of the types of teams. Overall it was a very effective model for both types but did come with some issues especially high log losses. For comparison, I also put the data through a Random Forest model to further study the impact of these features. While these models also came with high log losses, they showed high amounts of accuracy and were very informative about the predictor variables in the data.

For the Logistic Regression model, I set the variable "X" to be all of the features in the data that were continuous numbers, meaning I left out data such as wins, losses, week number, etc. I also decided to leave out points scored for either team because that is too predictive; clearly the more points a team scores the more likely they are to win. The "y" variable was either Road Win or Home Win depending on which version of the model I was using. The data was then split into training and testing sets with the test size being about 20% of the data. In order to avoid overfitting, I put the training data through a GridSearchCV with 5-fold cross validation that included the Logistic Regression model as well as parameter "C" that used the logspace method from NumPy. I fitted the training data and was able to find the best score and best parameter as well as find R squared and the root mean squared error. Lastly, I used the predict method on the testing data for "X" and got the accuracy score from that prediction compared with the testing data for "y."

For Road Win, I was able to get an accuracy score and AUC score around 96%, which is very high. After running a Confusion Matrix, 133 out of 138 labels were correctly classified and a Classification Report showed the average precision and recall were both at 96%. When it came to the Home Win model, the accuracy score and AUC score were around 91% which is also high, but not quite as good as the Road Win model. For the Confusion Matrix and Classification Report, 126 out of 138 labels were correctly classified along with average precision and recall equal to the accuracy score. While the Home Win model was not as strong, it still showed very high accuracy when it came to predicting winning.

To dig deeper, I wanted to see the features that had the highest predictive power. In order to do that, I created another Logistic Regression model for each team with C set to the best parameter for each model. After fitting the training data, I found their coefficients and set up a pandas DataFrame that sorted each feature by its impact on the model. When it came to the values, the strongest could be either positive or negative; it was about the overall impact on the model. Each showed essentially the same values just with the negative ones being positive for the other team and vice versa. The most common types of statistics at the top were scoring plays (like touchdowns and field goals) and turnovers while statistics near the bottom included all types of yards (rushing, passing and total) as well as time of possession and special teams' plays. Much of this was as expected but really emphasized how unimportant yards are. According to this model, winning is more about scoring and turnovers then just racking up as many yards as possible no matter which type of offense a team is running. And based on these features, it makes sense which games were mislabeled as they tended to have higher amounts of touchdowns and field goals made.

As I mentioned earlier, a Random Forest model was also used and took less time to set up. Because of its nature, I did not need to run GridSearchCV but set my number of estimators to 1,000 and fit the training data. For the Road team version, I received an accuracy score as well as AUC score around 92%, 127 out of 138 labels correctly classified and the same average precision and recall as the accuracy score. For the Home team model, the accuracy score was around 86%, AUC score around 85%, 118 out of 138 labels correctly classified as well as an average precision around 85% and an average recall around 86%. So, yet again, the Home team model was weaker than the Road team model. And while both showed significantly high accuracy, the Random Forest models were not as strong as those from Logistic Regression.

Feature importance was much simpler to display this time around and also all of the values were positive. While the two teams had similar results, they were a bit different from the features highlighted by Logistic Regression. Some of the more common, stronger features were kickoffs, rushing attempts, total touchdowns and time of possession while weaker features included scores off turnovers, safeties and special teams' scores. This model seemed to emphasize scoring but in a slightly different manner. While touchdowns were important, kicking the ball off was important too which has to do with the amount of times a team scores plus rushing attempts and time of possession emphasize controlling the ball. The weaker features were more about infrequent scoring plays that came from outside of the offense. Once again, the importance of these features led to the mislabeled data as higher amounts of kickoffs and rushing attempts were common in the incorrect labeling. This of course was very different from the Logistic Regression features that focused more on all scoring plays and turnovers with less focus on how a team went down the field. However, both showed less inclination towards the impact of special teams.

Between the two models, there is a very high amount of accuracy, which is a good thing for many situations. However, with large log losses, you can be punished greatly for being confident and wrong. And as seen by the features, each model focused on different predictive variables, which is part of what led to different scores. There is definitely more preprocessing that could be done to help fix some of these issues but overall, because of their high accuracy scores, both models would do a good job of taking the statistics inputted and predicting whether the Road team or Home team would win, which was the main focus of the project. And together they indicate the different ways a team wins with one more focused on overall scoring and avoiding turnovers while the other is more about how a team scores along with added preference placed on running as well as controlling the football. They both perfectly symbolize the different coaching styles seen throughout the NFL.