National Football League Capstone Project

Introduction

The National Football League (NFL) is the biggest football league in the United States. The league brings a huge viewership, especially in the championship game, the Super Bowl. With each team fighting for a win, for a chance to win their division and eventually the Superbowl. The NFL provided Kaggle with an extensive dataset that captures data from every game in the week. The dataset provided will be used to figure what type of formation a team, specifically the Ravens, use to increase their expected points added from each play.

Data

The data was collected by the NFL and given to Kaggle. The complete data encompassed the whole regular season (10 weeks) worth of data, data on each player, and data about the game. Since the complete dataset is over 500,000 data, only week 2 will be merged with the main datasets: play data, player data, game data, and tackles data. Week 2 was chosen, because it was not too late or early in the season. So that teams could get accustomed to the new season.After that is done, then 500 data will be randomly selected to form the data frame that the model will be fitted with. The next step is cleaning the data.

The data was then cleaned using various methods. The first method was by removing unnecessary data. Some data, like player name, college, birth date, etc. were removed because they were all string data. Since those did not relate to the hypothesis that was being answered they were removed. The next method was replacing all null values with the average of the column. This was done as a way to keep the data so that it could be used in the modeling stage. Lastly, dummy variables were created from the column, 'formation'. This was so that formation could be used in the modeling stage to see if formation did have some relationship with expected points added.

The next step was splitting the data into a 70 training and 30 testing split. After the data was split into training and testing variables, different models were run to see which fitted the data the best. The mean square error and the mean absolute error were used to see how well the testing data predictions were. In the first model, a simple Standard Scaler was used. The next model used Simple Imputer with the method as the median, Standard Scaler, and a Linear Regression. The third model included the Simple Imputer with the method as the median, Standard Scaler, Select best K, and a Linear Regression The fourth and last model used Simple Imputer, Standard Scaler, and a Random Forest Regressor. After testing each model, the fourth model the best model was the Simple Imputer with the method as the mean, Standard Scaler, and Random Forest Regressor with the n_estimator as 112. The best model will be used to fit the data and predict how changing certain features would affect the expected points added. While running the last pipeline, the top four features were play result, yards to go, down, and pre-penalty play result. So those 4 features were toyed with in the modeling stage.

During the modeling stage there were 3 scenarios that were tested out. The scenarios are:

- Changing the play results to see how it would affect expected points added
- Play ran gave us a 4 yard gain, and only 6 more yards to go at a second yard.

- The play ran gave us a 14 yard gain (lost 2 yards due to a foul), and only have 10 more yards to go at a first yard

For scenario 2 and 3, different formations were used to see how that would affect the expected points added. For both of those scenarios, the best formation was the shotgun formation.

Conclusions and Recommendations

After doing extensive work, the best model found out that the shotgun formation was the best. It should be remembered that this only looks at one week's worth of data, and not the Raven's complete season. So it is not a fair assumption that the Ravens should use this formation every time. If the other week's data were to confirm the finding in this project, then I would caution always using this formation. If the Ravens were to use this formation all the time, then their play would become predictable. Which would make the expected points from that play lower than what was observed. There are several recommendations. The first recommendation would be to use shotgun formation in situations where it is dire to score. The next recommendation would be to run this experiment, but this time focus on what is the strongest offensive team that could be built to increase the expected points added from each play. The last recommendation would be to try running this experiment with all the dataset that were given, to see how that would change the expected points added.