Michael Pallante
MSDS 456
Assignment 4

For this assignment, we will be studying National Football League data containing regular season player statistics from the years 2012 through 2016. Our focus will be exclusively on offensive player statistics. The offensive categories included in our data are passing, rushing, and receiving statistics from every regular season game of each year from 2012 to 2016. The objective of this assignment is to use these NFL offensive player statistics as measurables to build a predictive model that predicts future regular season offensive touchdown output. We must also create a metric that accounts for all types of offensive touchdowns combined. This metric will be the response variable for our predictive model and will be further described shortly.

Before we can enter the modeling phase, we must prepare the data set for further use. We will be dividing the data into a training data set and a test data set to try to improve the accuracy of our modeling results. The training data set will be composed of our NFL offensive player statistics data from the years 2012 to 2016, and the test data set will only contain the data from the year of 2016. We will be building our models on the training data set and testing them on the test data set. To fulfill the objective of our assignment, we will be using an ensemble predictive model, which is one large evaluation model that consists of a series of component models. Each individual component model within our ensemble model will be a multiple linear regression model. We will also be developing a null model to use as a comparative measure to judge our model against. Using cross validation techniques, we will see whether or not our ensemble model performs better than the null model. The predictive performance of these models will be evaluated by the computation of their root mean-square error (RMSE) value, where the lower the RMSE value, the better fit and better performing the model is. We will compare our evaluation model with the null model to see if using previous season player statistics is a better predictor of future season player statistics than that of the basic null model. Next, we must take

a look at all of the variables at our disposal and further understand the data we are using. In Figure 1.0 below, we can see an overview of the type of data on hand in our training set.

Figure 1.0: Overview of Training Data Set

```
> str(training)
'data.frame':   54033 obs. of  15 variables:
 $ Season   : chr  "2012" "2012" "2012" "2012" ...
 $ Team     : chr  "NYG" "NYG" "NYG" "NYG" ...
 $ name     : chr  "D.Wilson" "D.Hixon" "A.Bradshaw" "E.Manning" ...
 $ pass.att : num  0 0 0 32 0 0 0 0 0 0 ...
 $ pass.comp: num  0 0 0 21 0 0 0 0 0 0 ...
 $ passyds  : num  0 0 0 213 0 0 0 0 0 0 ...
 $ pass.tds : num  0 0 0 1 0 0 0 0 0 0 ...
 $ rush.att : num  2 0 17 0 0 0 0 0 0 0 ...
 $ rushyds  : num  4 0 78 0 0 0 0 0 0 0 ...
 $ rushtds  : num  0 0 1 0 0 0 0 0 0 0 ...
 $ recept   : num  0 3 2 0 4 6 4 1 1 0 ...
 $ recyds   : num  0 55 15 0 38 58 40 1 6 0 ...
 $ rec.tds  : num  0 0 0 0 0 0 1 0 0 0 ...
 $ games    : num  1 1 1 1 1 1 1 1 1 1 ...
 $ total.tds: num  0 0 1 1 0 0 1 0 0 0 ...

> summary(training)
   Season             Team               name             pass.att       pass.comp         passyds
 Length:54033       Length:54033       Length:54033       Min.   : 0.000   Min.   : 0.0000   Min.   : -8.000
 Class :character   Class :character   Class :character   1st Qu.: 0.000   1st Qu.: 0.0000   1st Qu.:  0.000
 Mode  :character   Mode  :character   Mode  :character   Median : 0.000   Median : 0.0000   Median :  0.000
                                                          Mean   : 1.334   Mean   : 0.8267   Mean   :  9.561
                                                          3rd Qu.: 0.000   3rd Qu.: 0.0000   3rd Qu.:  0.000
                                                          Max.   :65.000   Max.   :43.0000   Max.   :527.000
   pass.tds          rush.att          rushyds          rushtds          recept          recyds          rec.tds
 Min.   :0.00000   Min.   : 0.000   Min.   :-18.000   Min.   :0.0000   Min.   : 0.0000   Min.   :-15.000   Min.   :0.00000
 1st Qu.:0.00000   1st Qu.: 0.000   1st Qu.:  0.000   1st Qu.:0.0000   1st Qu.: 0.0000   1st Qu.:  0.000   1st Qu.:0.00000
 Median :0.00000   Median : 0.000   Median :  0.000   Median :0.0000   Median : 0.0000   Median :  0.000   Median :0.00000
 Mean   :0.05939   Mean   : 1.017   Mean   :  4.251   Mean   :0.0288   Mean   : 0.8267   Mean   :  9.562   Mean   :0.05939
 3rd Qu.:0.00000   3rd Qu.: 0.000   3rd Qu.:  0.000   3rd Qu.:0.0000   3rd Qu.: 1.0000   3rd Qu.:  4.000   3rd Qu.:0.00000
 Max.   :7.00000   Max.   :38.000   Max.   :251.000   Max.   :4.0000   Max.   :18.0000   Max.   :329.000   Max.   :4.00000
     games      total.tds
 Min.   :1   Min.   :0.0000
 1st Qu.:1   1st Qu.:0.0000
 Median :1   Median :0.0000
 Mean   :1   Mean   :0.1476
 3rd Qu.:1   3rd Qu.:0.0000
 Max.   :1   Max.   :7.0000
```

The original data contained some defensive, kicking, and other special teams statistics. We dropped those statistics from the data set, as we only needed offensive statistics for our analysis. The 15 variables within out training data set above are of the main focus and will be touched on further when we discuss the component models of our evaluation model. In short though, our training set contains basic offensive statistics and is outlined by player, by team, by game, and by season. As referenced earlier, we needed to create a metric that accounted for all types of offensive touchdowns combined. We developed a variable as a summary measure for total touchdowns. This measure was calculated within our data sets and then added to both of our training and test data sets. The metric was named

total.tds, which is the sum of passing touchdowns, rushing touchdowns, and receiving touchdowns by a player in a game. In other words, total touchdowns are the total number of offensive touchdowns by a player in a game. This will be used as our response variable throughout this assignment. Our predictor variables will be all numerical variables in our data set except for games, as that statistic is not needed for this purpose. The other character variables are not needed for this evaluation as well. It should also be noted that fortunately, there is no missing data within our data set and that imputation methods did not need to be used. This data is fully ready for modeling use.

To further elaborate on our modeling strategy, we will compute the average RMSE of the component models within our evaluation model through cross validation techniques, and then weigh it against the average RMSE of the null model. The null model predicts the mean response value of every player. This means that the null model predicts the mean touchdowns of every player. The null model is very basic, and we always want to outperform it. If not, our predictive model would be considered unsuccessful. The methodology for our component models that was deemed best was creating multiple linear regression models (lm) by offensive category using the specific statistics that fall under each category. This will give us unique models for each offensive category. The specifics of each model and each statistic used are seen in Figure 2.0 below, which gives a description of the component models within our evaluation model.

Figure 2.0: Overview of the Component Models within our Evaluation Model

1. **passing.model <- lm(total.tds ~ pass.att + pass.comp + passyds + pass.tds)**

   For this model, we will be predicting total touchdowns (total.tds) by specifically passing statistics only. These passing statistics include pass attempts (pass.att), pass completions (pass.comp), passing yards (passyds), and passing touchdowns (pass.tds). This should give us a strong idea of how much of an impact that passing statistics have on total touchdowns.

2. **rushing.model <- lm(total.tds ~ rush.att + rushyds + rushtds)**

   For this model, we will be predicting total touchdowns (total.tds) by specifically rushing statistics only. These rushing statistics include rush attempts (rush.att), rushing yards (rushyds), and

rushing touchdowns (rushtds). This should give us a strong idea of how much of an impact that rushing statistics have on total touchdowns.

3. **receiving.model <- lm(total.tds ~ recept + recyds + rec.tds)**

   For this model, we will be predicting total touchdowns (total.tds) by specifically receiving statistics only. These receiving statistics include receptions (recept), receiving yards (recyds), and receiving touchdowns (rec.tds). This should give us a strong idea of how much of an impact that receiving statistics have on total touchdowns.

4. **offensivetd.model <- lm(total.tds ~ pass.tds + rushtds + rec.tds)**

   For this model, we will be predicting total touchdowns (total.tds) by specifically offensive touchdown statistics only. These offensive touchdown statistics include passing touchdowns (pass.tds), rushing touchdowns (rushtds), and receiving touchdowns (rec.tds). This should give us a strong idea of how much of an impact that offensive touchdown statistics have on total touchdowns. This is likely the strongest indicator of total touchdowns out of all of our models because it considers scoring statistics only, and touchdowns are a scoring statistic.

Now that we have identified the component models that compose the evaluation model, as well as our modeling strategies, we can now run the evaluation model. Remember, this model was run on the training data set. After running the evaluation model, we then predict the entire evaluation model and its component models on the test data set. Once that is completed and cross validation techniques are complete, we are able to judge the performance of our evaluation model against the null model. Again, the performance metric for these models is the computation of their average root mean-squared error (RMSE) values, and the lower the RMSE means the better fit and better performing the model is. In Figure 3.0 below, we can see the performance of both models in comparison to each other.

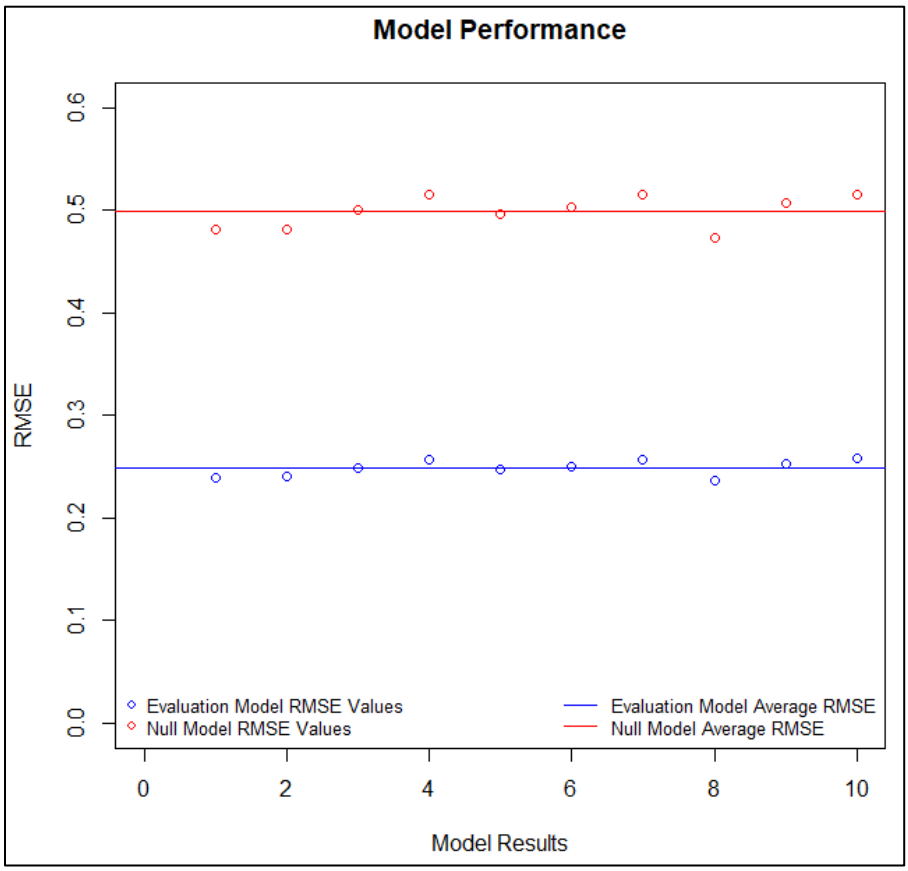
Figure 3.0: Model Performance Chart

Model	Average RMSE Value
Evaluation Model (Our Model)	0.248926
Null Model	0.499114

The results of the modeling and cross validation processes were very promising. Our evaluation model did outperform the null model and had a lower average RMSE value, proving to be a better fit. The margin in which the evaluation model outperformed the null model was significant, which is a great

sign. This means that our evaluation model predicts future regular season offensive touchdown output much better than the null model, which was the goal of this assignment. In Figure 4.0 below, we can see the performance of both models on a plot together, which can help us visualize the results better. Judging by our RMSE metric, both models actually performed really well, but the plot helps us see that our evaluation model was definitely a much better performer than the null model.

Figure 4.0: Model Performance Plot



Our model produced great results, however, we should still always reflect on the process to see how and where we could potentially improve performance. For example, our model uses very basic and traditional player statistics and does not take into account a lot of other statistics that we did not include in our evaluation. Our model also does not include a great wealth of data, as the sample size of 4-5 years could be too little, or it could even be too much if we need to project players who haven't

been in the league long. That leaves room for possible inconsistencies in our process. It is also hard to quantify some aspects of football, where evaluating with your eyes through the studying of game film is invaluable and often the best method of evaluation. If you see the skills of a player on tape that hasn't played often because he was a backup or is a rookie, you cannot necessarily project his statistics because he has not accumulated any statistics yet. The skills that jump out to you on that film, coupled with more playing time, can lead to that player having a breakout season that we simply cannot predict with numbers or past history. All of these details are ones that analysts should be conscious of during this process and try to account for before modeling.

This type of project can seemingly offer some value in multiple ways. For team front offices and from a player personnel and football operations perspective, this type of predictive model can be used to project future performance for players on their team. This can assist in player evaluation and decision making. A team may decide a player is not going to be productive enough to be worth keeping on their roster, or a team may be satisfied with a player and decides to keep the player. A team can also use such projections to look for potential trade targets from other teams. Perhaps a team could even use these predictions to get a baseline projection on league-wide win-loss records and performance. From a player's perspective, a player can possibly use this model to see where he ranks among the rest of the league and use his profile for contract leverage. From another perspective, fantasy football analysts can use these projections to project players' expected fantasy points production or to rank players in order of how they are projected to produce in a season. Overall, this predictive model has a variety of uses and can be a valuable resource.