Predicting the Success of Rookie NFL Kickers

Nikhil Ajjarapu

5/3/2020

Overview

In 2015, the NFL instituted a rule change regarding extra points in 2015 by moving them back to the 15 yard line, as they were becoming extremely routine and missing one was incredibly rare. In fact, according to the New York Times,the percentage of extra points made between 2000-2015 never dipped below 98%. Since 2015, however, the average has been hovering around ~94%, a clear and significant dropoff. In general, kicker mishaps have not only cost teams valuable points and killed drives, they've knocked entire teams out of the playoffs (see: the Minnesota Vikings). Given that the kicker position is critical to team success, it is apparent that being able to draft a good kicker is very important. However, a sole focus on kicker accuracy in college doesn't seem to be the magic formula to drafting successful NFL kickers. Thus, the focus of this project is to try and create a model using various statistical techniques that potentially could be more accurate in deciding which kickers would be successful.

Data Collection

Evaluation Mechanism

The first step was to decide how to quantify a kicker's success. The trivial choice was using NFL FG%. However, this stat was a bit reductionist relative to a kicker's success because it didn't show the distance of the kicks the kicker was attempting, and it ignored the concept of extra points completely. The next choice of evalutating a kicker's success was Weighted Career AV, which stands for Approximate Value. A measurement developed by PFR (Pro Football Reference, a famous football statistics website) founder Doug Drinen, it attempts to quantify single seasons by any player (methdology can be read here) and put a value over a player's entire career. However, in Drinen's own words, "If one player is a 16 and another is a 14, we can't be very confident that the 16AV player actually had a better season than the 14AV player" (taken directly from the methodology website). It seemed that the statistic was only relevant when comparing aggregate groups of players as opposed to individual ones. Thus, I decided to create my own statistic: Fantasy Points per Real Points (FP/RP). In fantasy football, scoring for kickers may vary from site to site, but it usually works as follows: extra points (XPs) receive one point, FGs from 1-50 feet receive 3 points, and 50+ yard FGs receive 5 points. Finally, every missed FG is -1 points. Thus, the overall formula to calculate this statistic was: **(total fantasy points over career) / (3 * Extra Points kicked + FGs kicked)**. There were several reasons for using this statistic:

1. The primary advantage of this statistic is that it controls for distance: it rewards FGs from longer distance, and since kicking is a low sample activity (only 16 games per season), every FG missed or made matters. Most statistics don't have a built-in mechanism for adjusting for distance, which can be very important in determining how good a kicker actually is.

2. FG% also suffers from increased sensitivity, as one missed kick can have a drastic effect on the percentage value, which FP/RP adjusts for by using a points system and only subtracting one point for every missed FG.
3. Using volume stats (total points scored) was also in consideration, but this would unfairly punish newer kickers that were a lot more accurate than older kickers but have had much shorter careers, as well as talented kickers whose careers were cut short due to injuries. To adjust for this, FP/RP is divided by the kicker's real points scored to adjust for volume.

Most FP/RP values ranged between 0.6 - 1.1, and from inspection of the dataset compiled, it was a lot more accurate than most other traditional statistics. For example, it ranked Justin Tucker very highly and Roberto Aguayo very low, even though Roberto Aguayo was one of the most accurate college kickers in history and was ranked very highly by draft experts (and went on to be one of the biggest busts in the history of the NFL). In general, it ranked successful NFL kickers with long careers highly, which made it the ideal evaluation functionn.

To collect these features, Python and the modules `requests` and `BeautifulSoup` were used to scrape the Pro Football Reference NFL website and compile the data into a CSV file for R to read. One major caveat was that only kickers who have played in at least 16 games (one season's worth) from 2000 - 2019 were included. This was to avoid small sample sizes (a kicker having 1.2 FP/RP on 5 kicks total), but also the website used to compile college data only supported kickers from 2000 onwards. All the code used to compile the data is available in the Github repository associated with this project.

Feature Selection

To represent each kicker in college, the features used were: total FGA (field goals attempted), total FGM (field goals made), total XPA, and total XPM. In addition, to measure improvement/decline from season to season, average change in FG% from season to season, as well as average change in XP% from season to season were included. Both these values were set at 0 if the kicker had only played for one season at their college (as there was no improvement or decline). Unfortunately, the dataset was severely limited by the amount of data available for free and easily scrapable, which is why the list of features is fairly small. Similar to above, Python and the modules `requests` and `BeautifulSoup` were used to scrape the Pro Football Reference CFB (college football) website and compile it into a CSV file.

Model Fitting

In order to test which model would be best for this activity, 7 unique models were built and tested: two linear models, a KNN regression model, a lasso regression and ridge regression model, and finally two logistic regression models. Using various models as well as different sets of features for certain models allows us to create the most accurate predictor for FP/RP. The performance of every model was measured using RMSE (root mean squared error) to determine which had the least error, with smaller RMSE values indicating higher accuracy. In order to reduce Monte Carlo variability, 200 repeated random samples of the data for each model were used to find the true RMSE values.

. . . .

Model 1: Linear Model

The first model built was the basic linear model, with the `lm` command. While lasso and ridge regression are superior to linear models, the linear model was included as well as a benchmark for the other models. There are two models: one with interactions and one without, to see which performs better. As the data isn't plottable on a 2D graph (multiple X values), the actual vs predicted values were graphed, with a y = x line to indicate perfect fit.

```
LINEAR REGRESSION MODEL (without interactions) - RMSE: 0.08100165
```

```
LINEAR REGRESSION MODEL (with interactions) - RMSE: 0.1066613
```

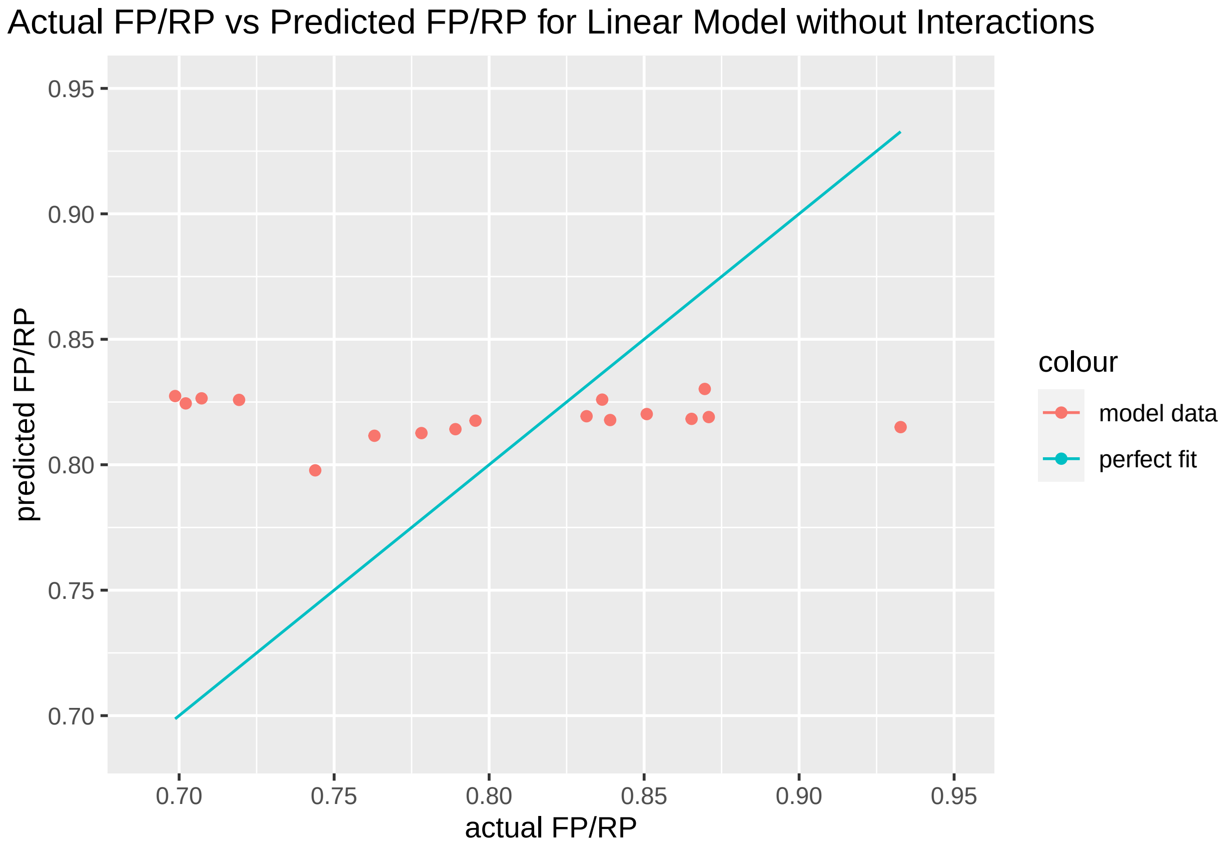

Figure 1: Measuring model performance by plotting model output vs a ideal perfect model. The linear model is not too strong as it predicts too many similar values.

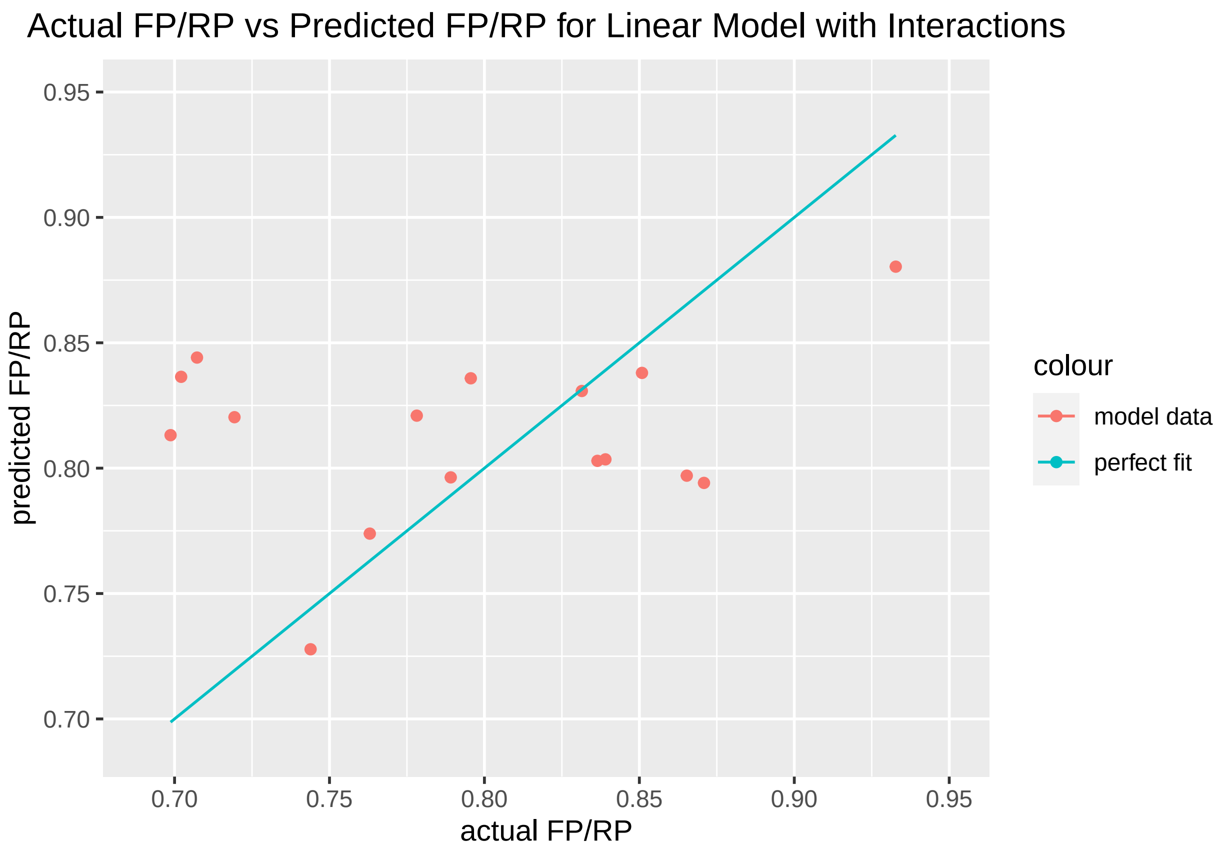

Figure 2: The linear model without interactions is even weaker and has higher variability in its predictions.

Model 2: kNN Model

The second model tested was a k Nearest Neighbors (kNN) model. Due to dataset limitations, only k-values up to 68 could be tested, even though ideally there would be more samples and the k-value would be higher. Ultimately, the optimal k was 60, with a RMSE value of ~0.07. Below, the k vs RMSE graph was included, which represents the accuracy of the kNN model at different values of k, from which we can see the optimal k graphically.

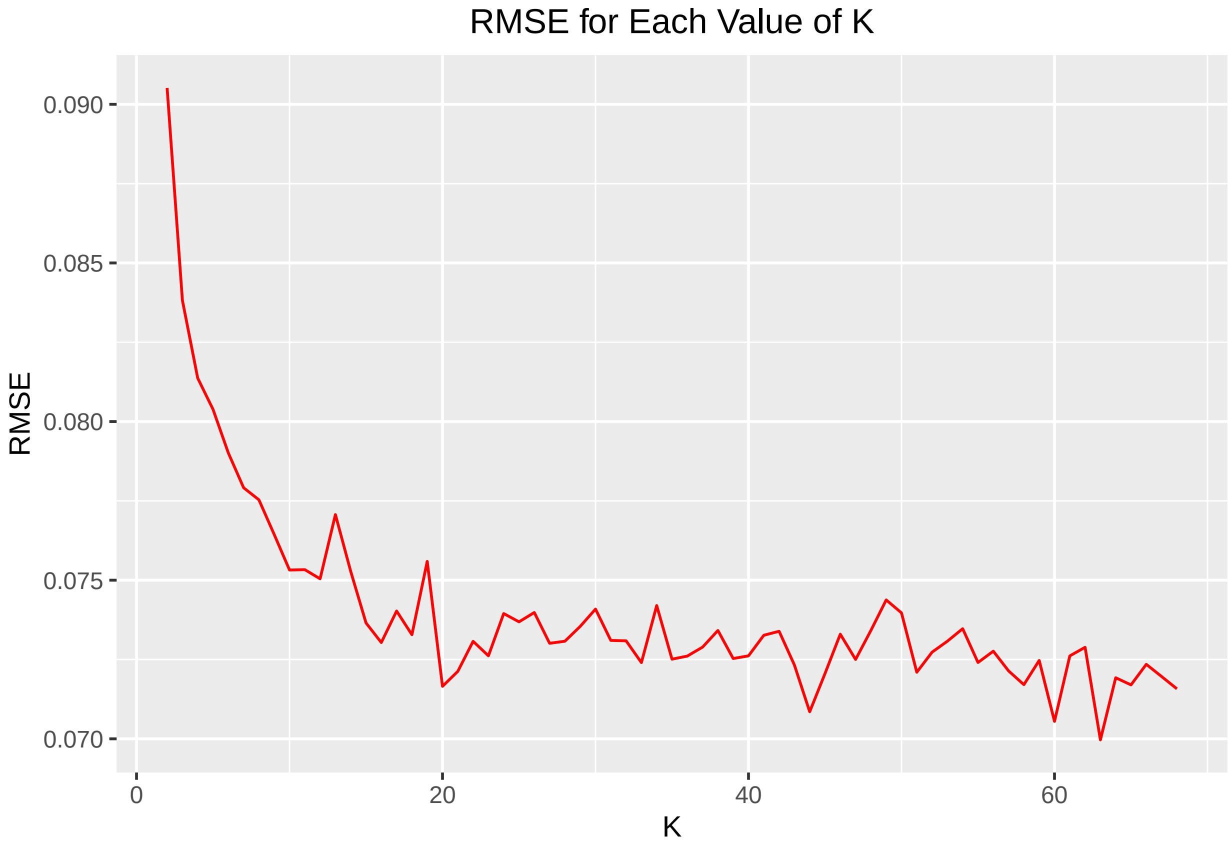

Figure 3: kNN vs RMSE plot. Helps show which k value should be used.

```
KNN ( k = 63 ) - RMSE: 0.06996509
```

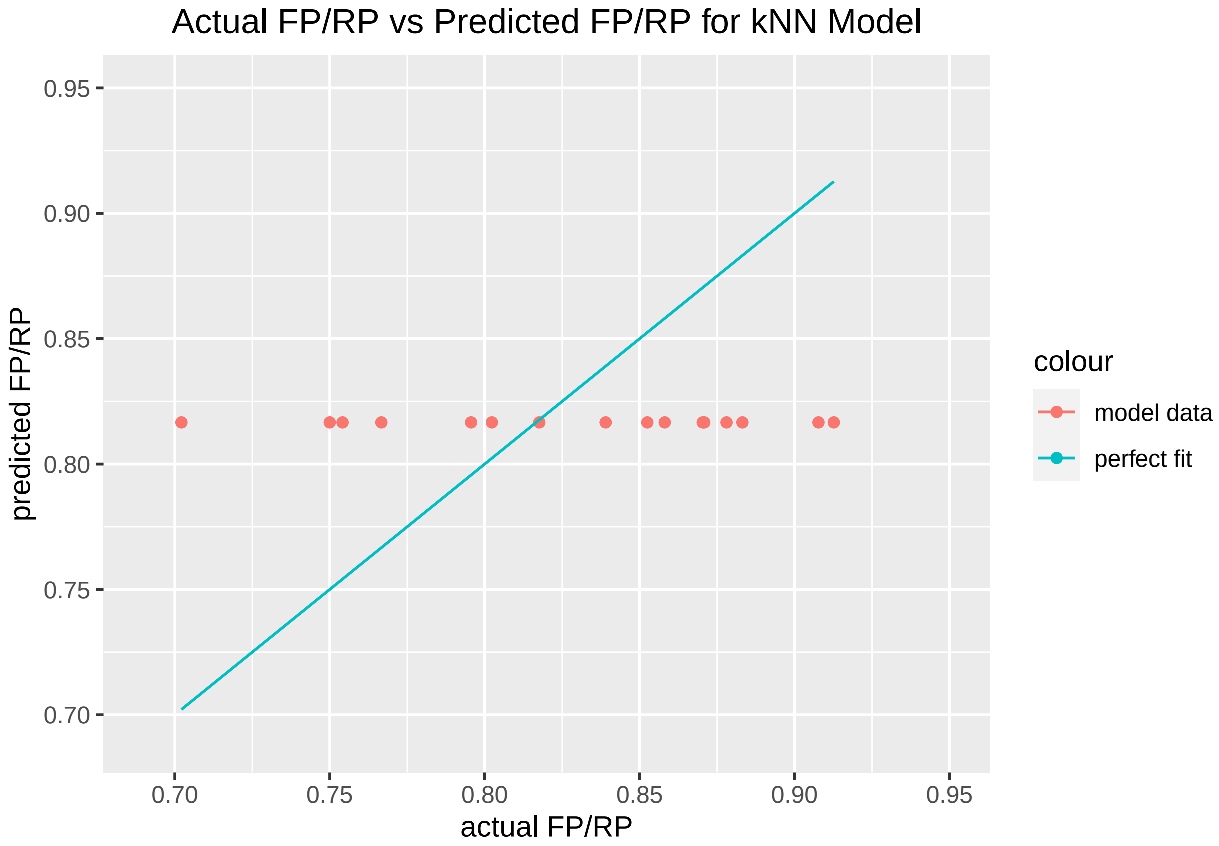

Figure 4: Data is mostly uniform, which is reflective of the bias in the data but also of the parameters of the model.

Model 3: Regularized Linear Regression

The next set of models were lasso and ridge regression. These supplant the vanilla linear models created previously.

```
LASSO REGRESSION - RMSE: 0.04510972
```

```
RIDGE REGRESSION - RMSE: 0.04510972
```

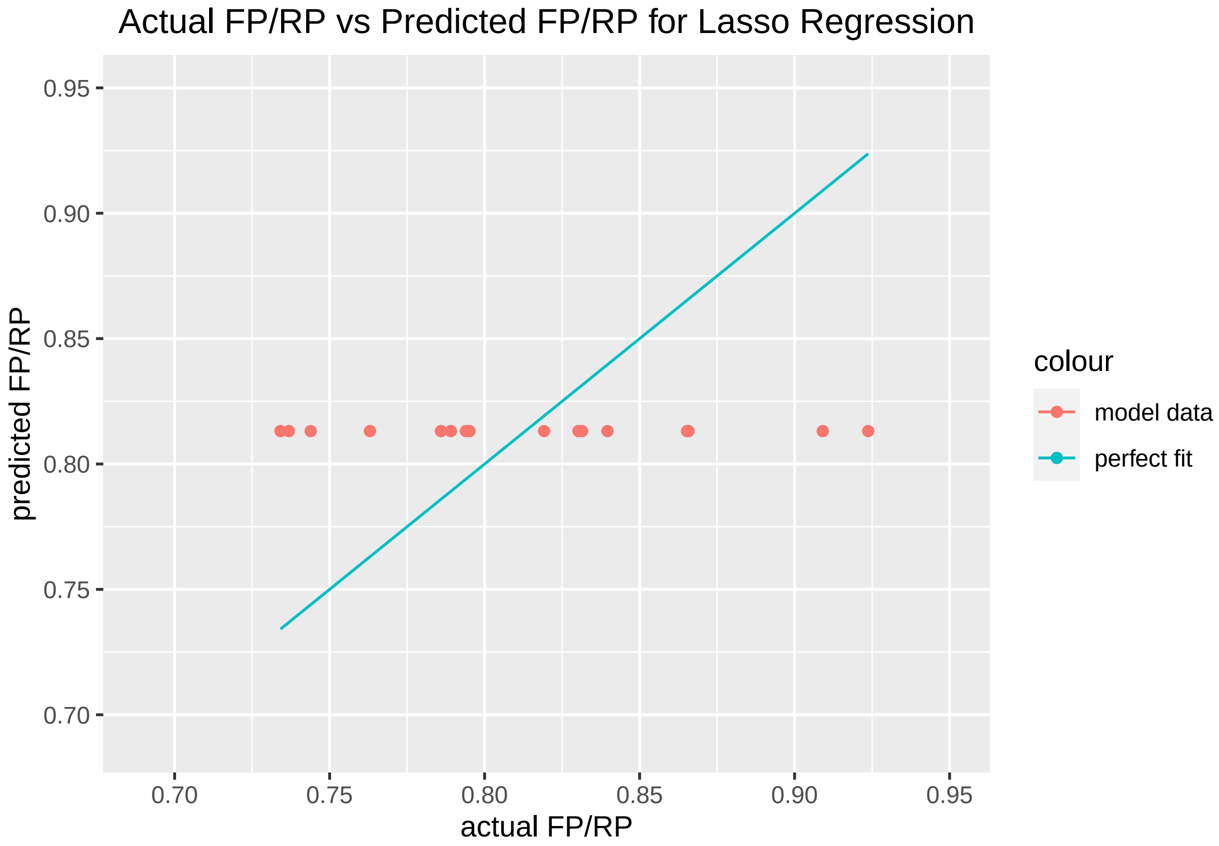

Figure 5: Although lasso regression looks similar to previous models to the human eye, the scale of the data is so small it is important to remember even a 0.01 difference in the estimate can make a significant difference in RMSE.

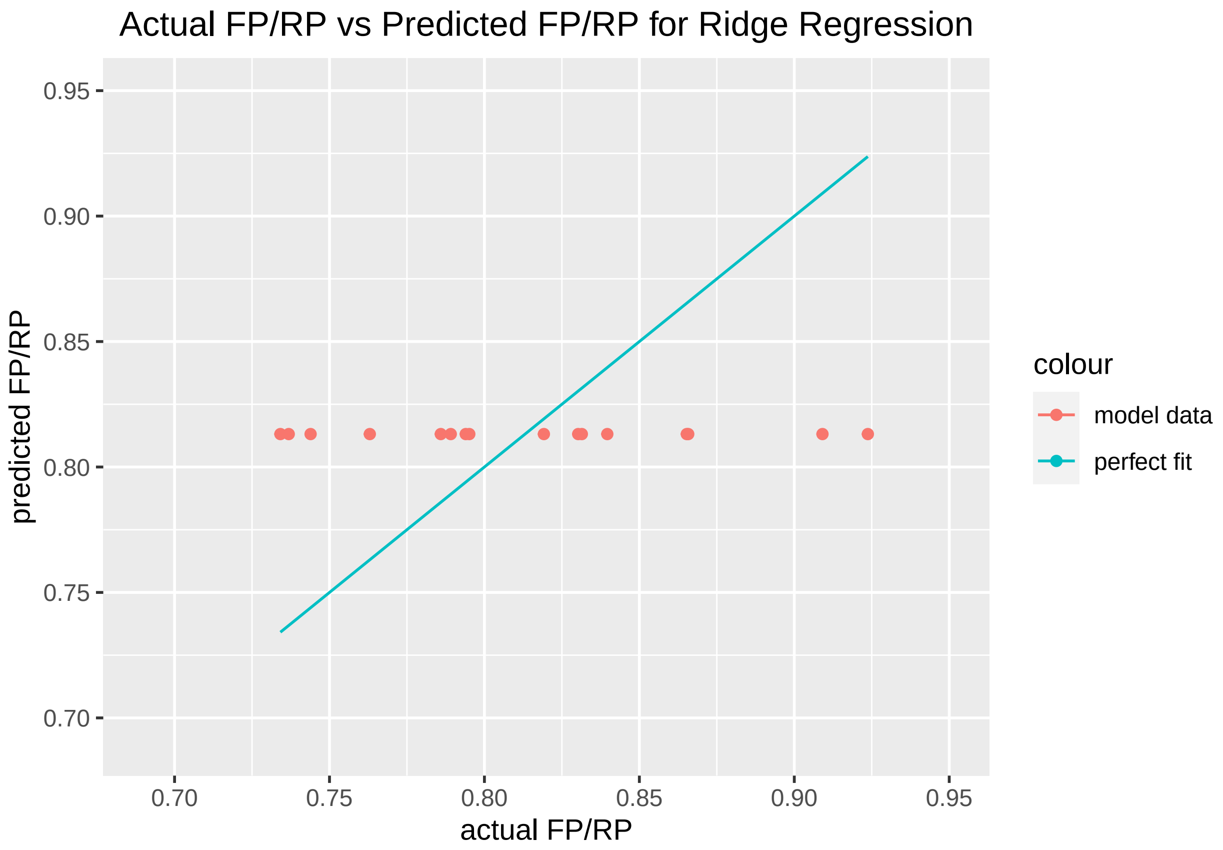

Figure 6: Although slightly less accurate, ridge regression is still a step above the other models even if it is not apparent.

Model 4: Logistic Regression

The final family of models that were tested were logistic regression models. Two unique models were included: one with interactions and one without, similar to before.

```
LOGISTIC REGRESSION (without interactions) - RMSE: 0.07971993
```

```
LOGISTIC REGRESSION (with interactions) - RMSE: 0.1077362
```

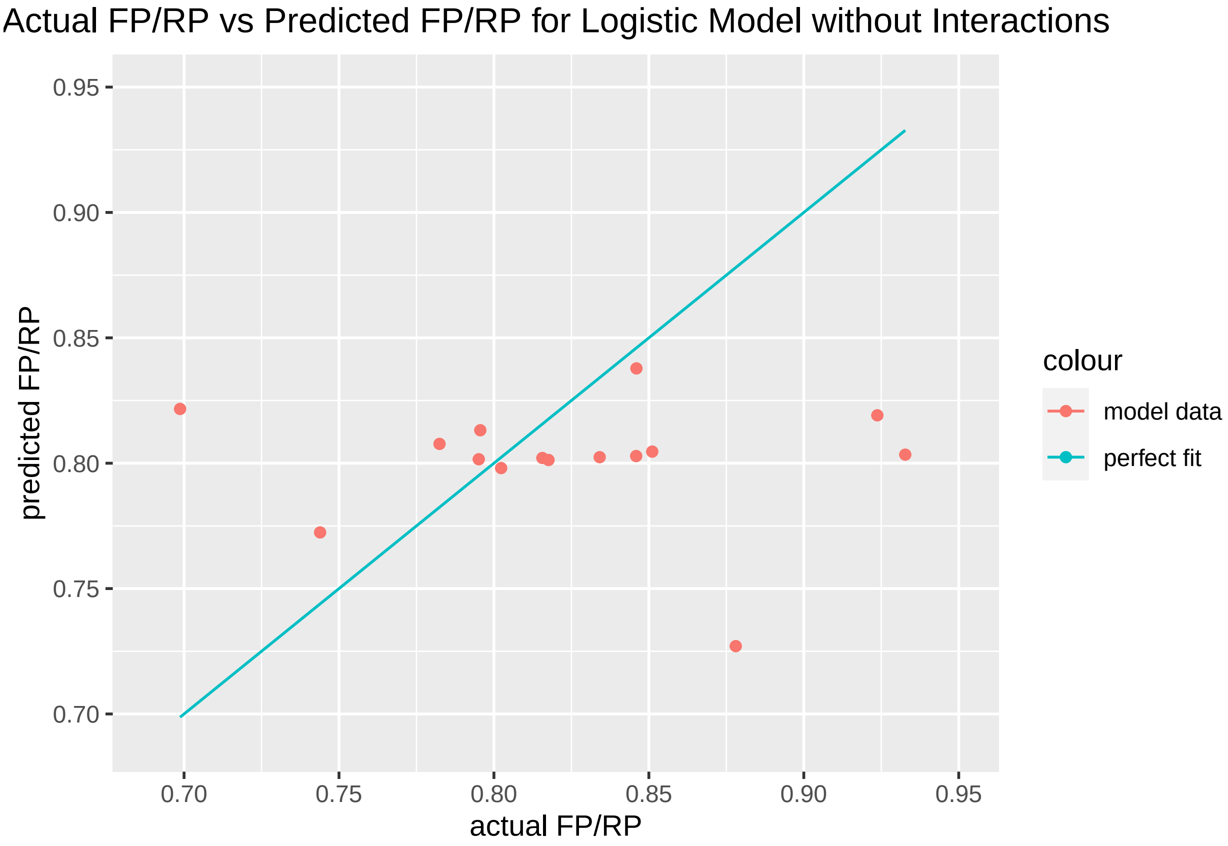

Figure 7: We can see that the logistic model without interactions, while more accurate than the other logistic model, still has high amounts of variability.

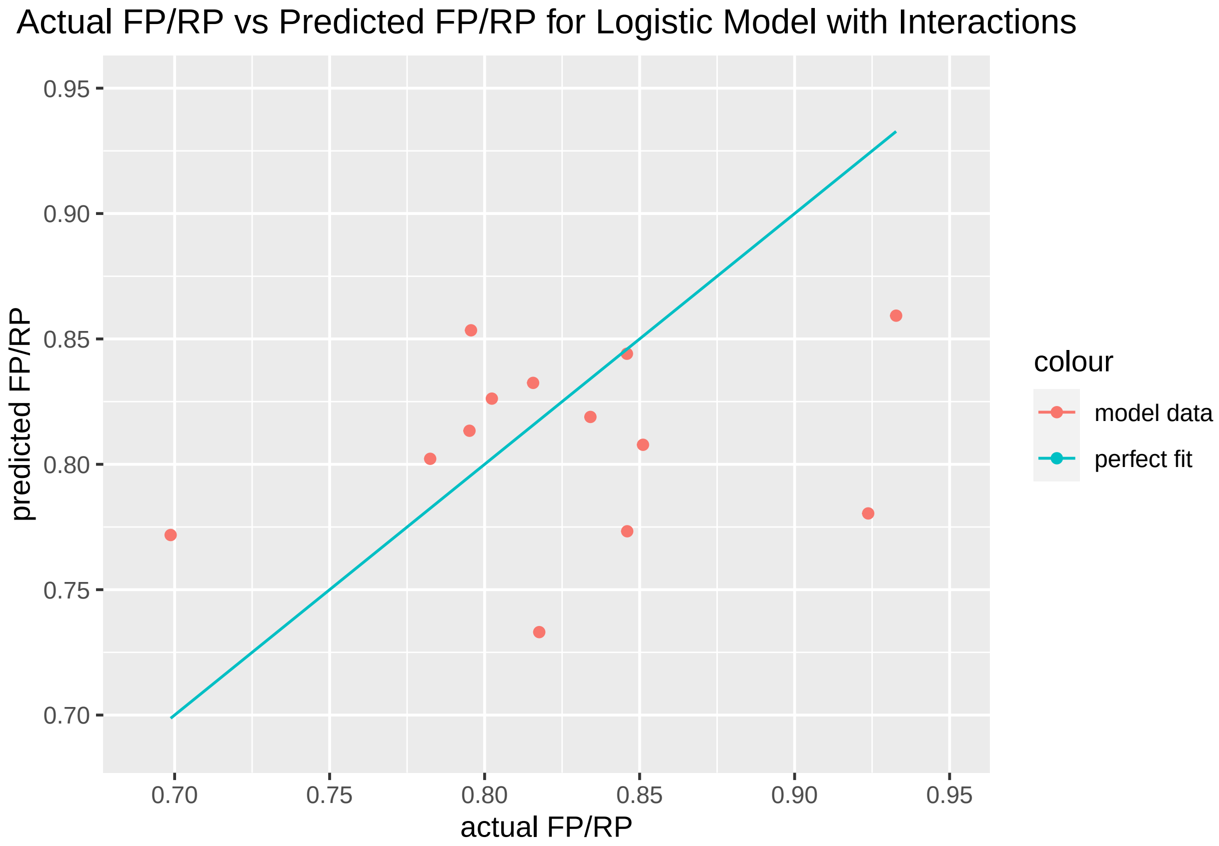

Figure 8: The logistic model with interactions, which is clearly not very reliable as it has (relatively) high amounts of error.

Results and Conclusion

Each model had 200 iterations at 80%/20% train/test splits. This was to be able to run cross-validation and get an accurate estimate of RMSE error. `cv.gamlr()` offered its own cross-validation, but it did not seem applicable to this situation as it calculated RMSE values to determine the optimal lambda, not for all models.

Of all the models, Lasso and Ridge Regression seem to be performing the best, at an approximate RMSE value of ~0.04. As the average FP/RP value is ~0.80, this means the regularized linear regression models can predict FP/RP values within ~±5%. This is in line with the idea that regularized linear models would perform better than regular linear models, as both lasso and ridge regression both surpass linear regression in performance in this case. kNN and logistic regression were the two wild cards, and both turned out to be lower in accuracy as well. kNN does not follow a strict path but rather relies on surrounding data, which turned out to be not as useful in this scenario. Similarly, logistic regression did not seem to be as applicable in this scenario, which indicates that the kicker data follows a somewhat linear path (in higher dimensions).

Ultimately, this means that it is possible to accurately predict the value of a kicker based on various college-level statistics. While the small sample sizes might be cause for hesitation, such a model definitely merits some attention. One thing to note, however, is the higher performing models are typically predicting the same value for each kicker. This is a cause for concern, as while it minimizes the error, it does not really match what happens in real life. However, this can be resolved with larger sample sizes, as the model would be able to have more varied predictions if it could have data of kickers that have not made it to the NFL either. Unfortunately, larger amounts of data are usually not available to the public or are behind paywalls, which limits the capacity of the project. Some additional questions to consider would be to see if NFL or college coaching had an measurable, statistical effect on kicking, as well as seeing if certain colleges are better at creating quality NFL kickers than others. Ideally, models such as the ones presented in this report will one day become standard so fans don't have to watch a playoff game or even a Super Bowl slip away from a missed kick (see: the Buffalo Bills and the Chicago Bears).