Outliers and Overperformance: A Data-Driven Analysis of NFL Seasons

AJ Fong

July 2025

Data Collection

For the data collection, I created a Python function that used the Selenium package to scrape live data from the website profootball-reference.com for NFL team stats from 1965-2025. I used this function to create .txt files that I then uploaded into R. With these files, I merged and cleaned the data based off of the year and team names. This allowed for smooth and efficient data extraction and cleaning.

```
##   Year                Tm  W L Win_pct Rk  G  Pf Total Yds Total Ply Y/P TO FL
## 1 1965   Baltimore Colts 10 3   0.769  3 14 389      4598       887 5.2 36 19
## 2 1965     Chicago Bears  9 5   0.643  2 14 409      4897       870 5.6 28 16
## 3 1965  Cleveland Browns 11 3   0.786  5 14 363      4398       836 5.3 25  9
## 4 1965    Dallas Cowboys  7 7   0.500  7 14 325      3995       833 4.8 35 17
## 5 1965     Detroit Lions  6 7   0.462 12 14 257      3303       853 3.9 41 15
## 6 1965 Green Bay Packers 10 3   0.769  8 14 316      3601       781 4.6 26 12
```

Creating & Optimizing Linear Models

I initially placed my data into the base R lm function with the response variable being win percentage to create the first model. I then used the summary function to check the significance level of each predictor as well as the R^2 and adjusted-R^2 values to help determine the fit of the model. I proceeded to use multiple different techniques such as BIC/AIC, VIF, InvResPlot, and BoxCox to optimize and generate the best model. The BIC/AIC methods help in deterimining the optimal amount of predictors through their model selection criteria. VIF is used to determine collinearity between different variables. It tells you how much the variance of a regression coefficient is inflated due to collinearity with other predictors. InvResPlot and BoxCox show whether or not a transformation (log, square root, etc) are useful in helping to more effectively predict the response variable. With these models, I checked the residual plots to determine if all normality conditions are met inside of the model to prevent biasness.

Note: Due to the missing values inside of data from 1968-1999, only 2000-2025 data was used in creating the models.

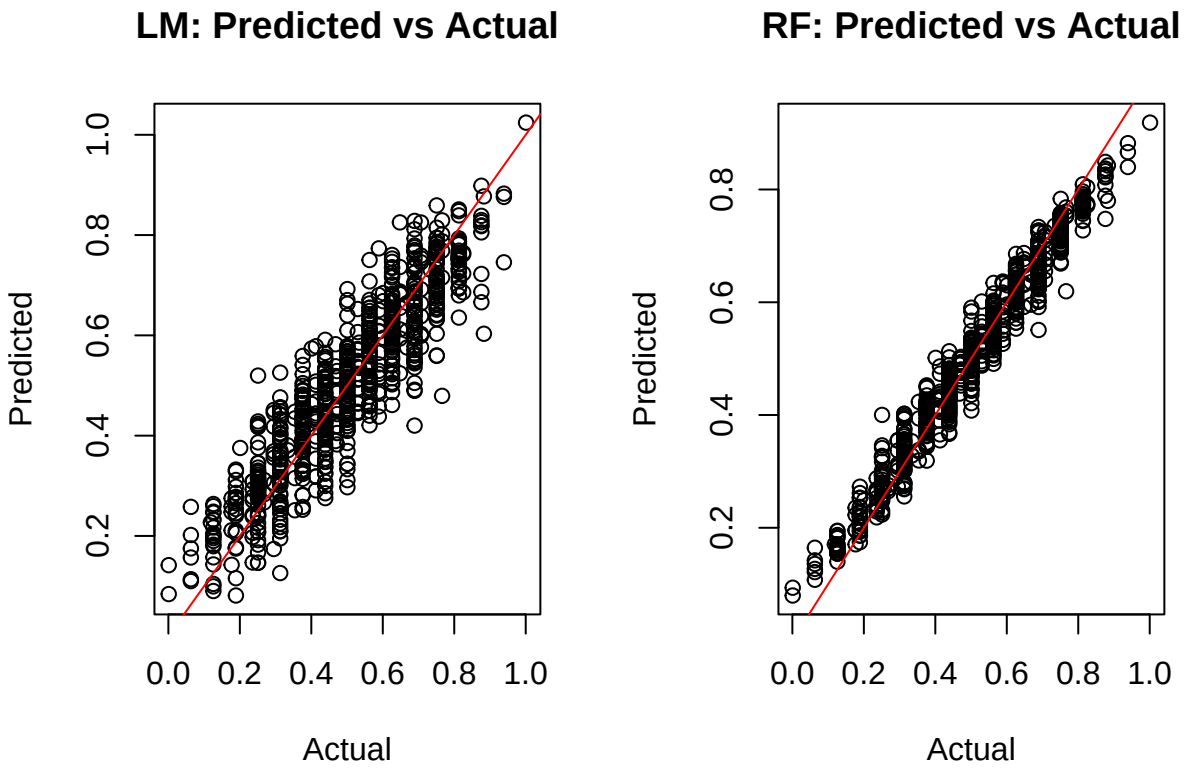
Another method I used to create a model was the random forest method. Random forest models build many decision trees and combines their results to make a more accurate model. It uses bootstrap sampling and random selection to prevent overfitting and add diversity.

Choice of Model

```
##
## Call:
```

```
## lm(formula = Win_pct ~ TO + `Pass TD` + `NY/A` + `Rush TD` +
##     `Pen Yds` + `Sc%` + Cmp_D + `Pass Att_D` + `Pass TD_D` +
##     `NY/A_D` + `Rush Att_D` + Pen_D + `Sc%_D` + `TO%_D`, data = total_stats_cleaned)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.271061 -0.053442 -0.002255  0.055499  0.284377
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.933e-01  8.991e-02   6.598 7.34e-11 ***
## TO           -3.640e-03  5.091e-04  -7.150 1.88e-12 ***
## `Pass TD`     3.080e-03  7.049e-04   4.370 1.40e-05 ***
## `NY/A`        2.345e-02  6.643e-03   3.530 0.000438 ***
## `Rush TD`     3.335e-03  7.727e-04   4.316 1.78e-05 ***
## `Pen Yds`    -8.858e-05  2.057e-05  -4.306 1.86e-05 ***
## `Sc%`         8.587e-03  9.881e-04   8.690  < 2e-16 ***
## Cmp_D        -7.842e-04  1.788e-04  -4.385 1.30e-05 ***
## `Pass Att_D`  7.890e-04  1.478e-04   5.340 1.19e-07 ***
## `Pass TD_D`  -2.909e-03  6.975e-04  -4.171 3.34e-05 ***
## `NY/A_D`     -3.558e-02  7.532e-03  -4.724 2.71e-06 ***
## `Rush Att_D` -4.889e-04  8.839e-05  -5.531 4.25e-08 ***
## Pen_D         6.827e-04  2.126e-04   3.211 0.001374 **
## `Sc%_D`      -8.529e-03  1.019e-03  -8.369 2.40e-16 ***
## `TO%_D`       6.028e-03  1.057e-03   5.703 1.63e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07999 on 844 degrees of freedom
##   (855 observations deleted due to missingness)
## Multiple R-squared:  0.8301, Adjusted R-squared:  0.8273
## F-statistic: 294.5 on 14 and 844 DF,  p-value: < 2.2e-16

##           TO    `Pass TD`       `NY/A`    `Rush TD`    `Pen Yds`        `Sc%`
##     1.692669     3.713088     3.740815     2.155399     1.190026     6.233507
##        Cmp_D `Pass Att_D`  `Pass TD_D`     `NY/A_D` `Rush Att_D`        Pen_D
##     6.637506     6.054383     2.018720     2.804745     2.313913     1.158055
##      `Sc%_D`      `TO%_D`
##     4.519752     1.616724

##
## Call:
##  randomForest(formula = Win_pct ~ ScPct + RushAttD + ScPctD +      NYA + NYAD + TotalYds + TOPctD +
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 5
##
##           Mean of squared residuals: 0.008252171
##                     % Var explained: 77.61

##   intercept       RMSE  Rsquared        MAE      RMSESD RsquaredSD       MAESD
## 1      TRUE 0.08106616 0.8195133 0.06485105 0.003768785 0.02109842 0.002765133

##   mtry       RMSE  Rsquared        MAE      RMSESD RsquaredSD       MAESD
```

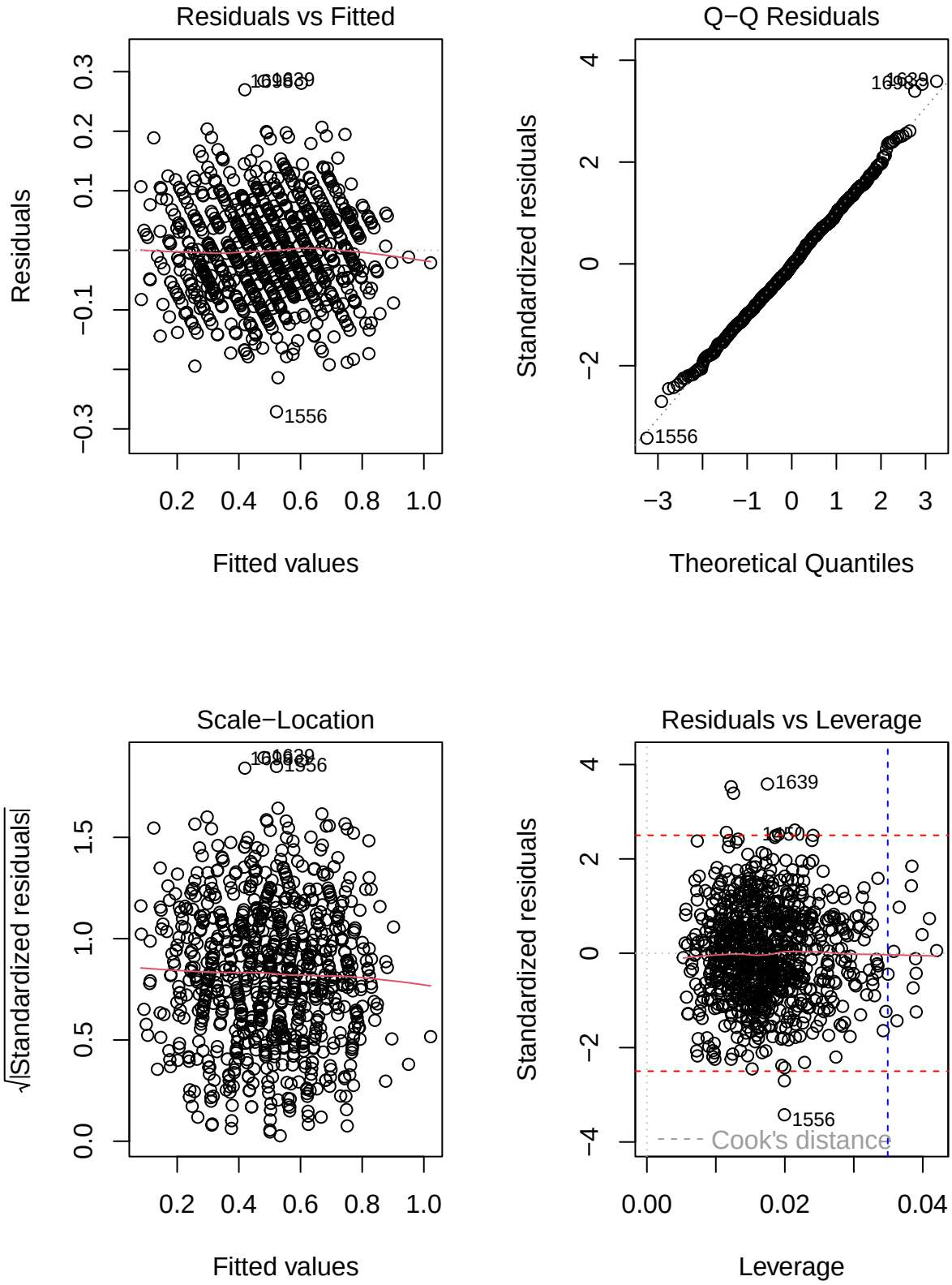
```
## 1    2 0.09329327 0.7855383 0.07496117 0.005504831 0.02859230 0.004843781
## 2    9 0.09204221 0.7830166 0.07334694 0.005943073 0.03685152 0.004864478
## 3   16 0.09285181 0.7763540 0.07356852 0.006176349 0.03968165 0.004654323

## Linear Model Results: 0.08971885 0.7943557 0.07075526

## Random Forest Results: 0.09047137 0.7902412 0.07169284
```



When selecting the best model, I evaluated several diagnostic metrics tailored to each modeling approach. For the linear regression model, I focused on the adjusted R^2 to assess model fit while accounting for complexity, p-values to determine the statistical significance of individual predictors, the Mean Squared Error (MSE) of the residuals to gauge prediction accuracy, and Variance Inflation Factors (VIFs) to detect potential multicollinearity.

For the random forest model, I examined the percentage of variance explained as a proxy for goodness of fit, the residual MSE to assess prediction performance, and variable importance scores to identify which predictors contributed most to the model.

Model Validity



The residual plot assesses whether the residuals exhibit constant variance. In this case, the red trend line is nearly flat without a noticeable fanning or curvature. This indicates the assumption of constant variance has been met. The Q-Q plot evaluates the normality of the residuals. Since the points closely follow the reference line, the residuals appear to be normally distributed. The scale-location plot checks for homoscedasticity, or equal spread of residuals. The red trend line is mostly flat with no indication of a pattern which supports the assumption of homoscedasticity. The residual vs leverage plot helps to identify influential points, including high leverage outliers and "bad" high leverage points that can disproportionately affect the model and distort the R^2 value. Bad high leverage points appear in the top right and bottom left quadrants of the graph. In this model, there are no bad high leverage points but there are outliers that will be further examined.

Team Outliers

The teams listed below were identified as being outliers based on the criterion of having standardized residuals greater than $^+_-$2.5 from the model's trend line:

```
##      Year                 Tm  W  L Win_pct Expected_Win_pct Exp_Wins Win_Diff
## 1044 2004    Atlanta Falcons 11  5   0.688            0.519    8.304    2.696
## 1216 2009 Indianapolis Colts 14  2   0.875            0.629   10.064    3.936
## 1276 2011     Denver Broncos  8  8   0.500            0.339    5.424    2.576
## 1312 2012 Indianapolis Colts 11  5   0.688            0.427    6.832    4.168
## 1453 2016 San Diego Chargers  5 11   0.312            0.565    9.040   -4.040
## 1556 2020    Atlanta Falcons  4 12   0.250            0.531    8.496   -4.496
## 1639 2022  Minnesota Vikings 13  4   0.765            0.489    8.313    4.687
## 1698 2024 Kansas City Chiefs 15  2   0.882            0.592   10.064    4.936
##      Outcome
## 1044  L NFCC
## 1216    L SB
## 1276  L AFCD
## 1312 L AFCWC
## 1453     4th
## 1556     4th
## 1639 L NFCWC
## 1698    L SB
```

Upon examining the outliers, a notable pattern emerged; teams have increasingly tended to over or underperform relative to the model's expectations in recent years. With the model using data from 1998 - 2024, we would expect that there would be one outlier every three years if there are a total of eight. However, there were three of the eight in the last four years.This suggests a potential shift in league dynamics or model limitations.

Additionally, teams that significantly overperformed according to the model did not regress in the postseason. For example, the 15-2 Chiefs were 4.94 wins above expected and still went on to make the Super Bowl. Overperformance may reflect real team strength or factors not captured by the model instead of simple variance or luck.

When reducing the criteria from $^+_-$ 2.5 standard residuals to 2.25, this trend looks to taper off:

```
outlier_data <- total_stats_m3[which(abs(total_stats_m3$StdRes) > 2.25), ]
outlier_data[,c(1:5, 61:63)]
```

```
##      Year                   Tm  W  L Win_pct Expected_Win_pct Exp_Wins Win_Diff
## 952  2001    Carolina Panthers  1 15   0.062            0.279    4.464   -3.464
## 963  2001       Miami Dolphins 11  5   0.688            0.512    8.192    2.808
```

```
## 1044 2004      Atlanta Falcons 11  5   0.688           0.519    8.304    2.696
## 1067 2004  Pittsburgh Steelers 15  1   0.938           0.717   11.472    3.528
## 1104 2005 Tampa Bay Buccaneers 11  5   0.688           0.475    7.600    3.400
## 1121 2006 Jacksonville Jaguars  8  8   0.500           0.712   11.392   -3.392
## 1137 2006     Tennessee Titans  8  8   0.500           0.302    4.832    3.168
## 1216 2009   Indianapolis Colts 14  2   0.875           0.629   10.064    3.936
## 1225 2009      Oakland Raiders  5 11   0.312           0.185    2.960    2.040
## 1260 2010   San Diego Chargers  9  7   0.562           0.748   11.968   -2.968
## 1276 2011       Denver Broncos  8  8   0.500           0.339    5.424    2.576
## 1312 2012   Indianapolis Colts 11  5   0.688           0.427    6.832    4.168
## 1404 2015       Denver Broncos 12  4   0.750           0.572    9.152    2.848
## 1450 2016      Oakland Raiders 12  4   0.750           0.543    8.688    3.312
## 1453 2016   San Diego Chargers  5 11   0.312           0.565    9.040   -4.040
## 1534 2019    Green Bay Packers 13  3   0.812           0.610    9.760    3.240
## 1556 2020      Atlanta Falcons  4 12   0.250           0.531    8.496   -4.496
## 1570 2020   Kansas City Chiefs 14  2   0.875           0.669   10.704    3.296
## 1608 2021 New England Patriots 10  7   0.588           0.778   13.226   -3.226
## 1639 2022    Minnesota Vikings 13  4   0.765           0.489    8.313    4.687
## 1698 2024   Kansas City Chiefs 15  2   0.882           0.592   10.064    4.936
```

Another interesting observation involves the wins total of these outlier teams. As shown above, teams often finish last in their division or make the playoffs. This aligns with the idea that teams that are farther away from the average amounts of wins for a season (typically 8 wins in a 16-game season or 8.5 in a 17-game season) are more likely to deviate in areas not captured by the model (i.e. special teams, advanced stats, etc).

Conclusion

This project explored the relationship between NFL team performance metrics and season win percentage through both linear progression and random forest models. By collecting and cleaning over 50 years of NFL data, I was able to construct predictive models that achieve high accuracy and met key statistical assumptions.

The linear model demonstrated strong explanatory power with an adjusted R^2 of .83, while the random forest model captured 77.7% of the variance and highlighted important nonlinear interactions among predictors. Both models offered valuable insights into which team statistics most significantly influence winning outcomes.

After examining the model's residuals, there were several teams that substantially over or underperformed relative to expectations. In recent seasons, such outliers have become more frequent. This indicates a potential shift in league dynamics, strategy innovations, or model limitations. An important factor to consider is teams that teams who overperformed in the regualr season did not regress in the postseason. This suggests that their success may reflect real strength not fully captured by traditional statistical inputs.

Overall, this analysis illustrates the utility of data-driven modeling in evaluating NFL team performance and identifying patterns in over or underachievement. Future improvements could include integrating advanced stats such as EPA, special teams metrics, or injury data to further enhance predictive power and capture hidden drivers of team success.