SAM ASSAF, GARRETT ATKINSON,
CARLO LOPEZ HERNANDEZ

# PREDICTING PREMIER LEAGUE WINS USING FIFA RATINGS 2015-2020

# TABLE OF CONTENTS



# DATASET 1

**"FIFA 20 Complete Player Dataset" (Kaggle)**

- Spans FIFA 15 to FIFA 20, covering the last six versions of the FIFA video game.
- Encompasses 15,409 distinct player entries.
  - Includes attributes such as Player positions, personal details (nationality, club, date of birth, wage, salary), Statistics related to attacking, skills, defense, mentality, goalkeeper skills, and more.
  - Maximum, median, and variance of various ratings/measurements aggregated for each position group within a club
  - The sum of work rate for each position within a club

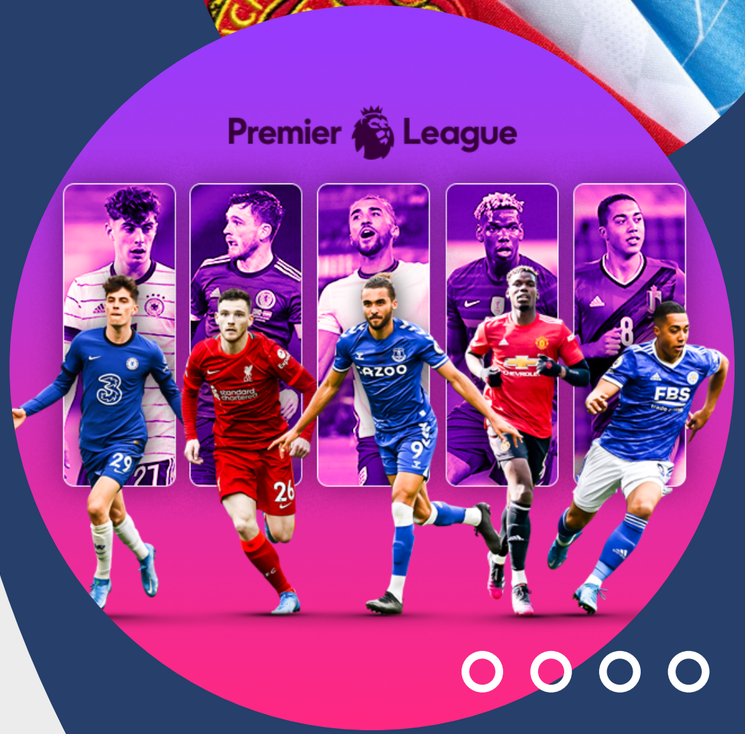
**Positional Categorization:**

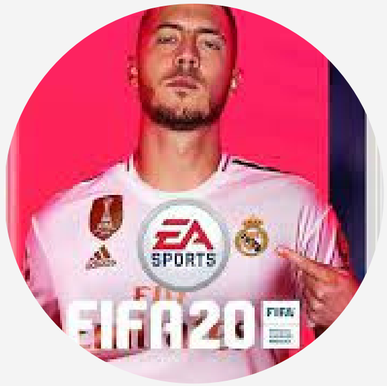
- **Attack:** Left Forward, Striker, Center Forward, Right Forward, Left Wing, Right Wing.
- **Midfield:** Left Midfielder, Center Midfielder, Central Attacking Midfielder, Central Defensive Midfielder, Right Midfielder.
- **Defense:** Left Wing-Back, Left Back, Center Back, Right Back, Right Wing-Back.
- **Goalkeeper:** Goalkeeper.

# DATASET 2

**"All Premier League Matches 2010-2021" (Kaggle)**

- “One of the most comprehensive datasets for the English Premier League”
- Encompasses information from:
  - 4,070 matches
  - over an 11-year period,
  - collected through web scraping,
  - featuring 113 distinct features.




## FIFA 20 DATA

- maximum, median, and variance of different aggregated ratings (across position groups), and maximum goalkeeper ratings.

+

## PREMIER LEAGUE DATA

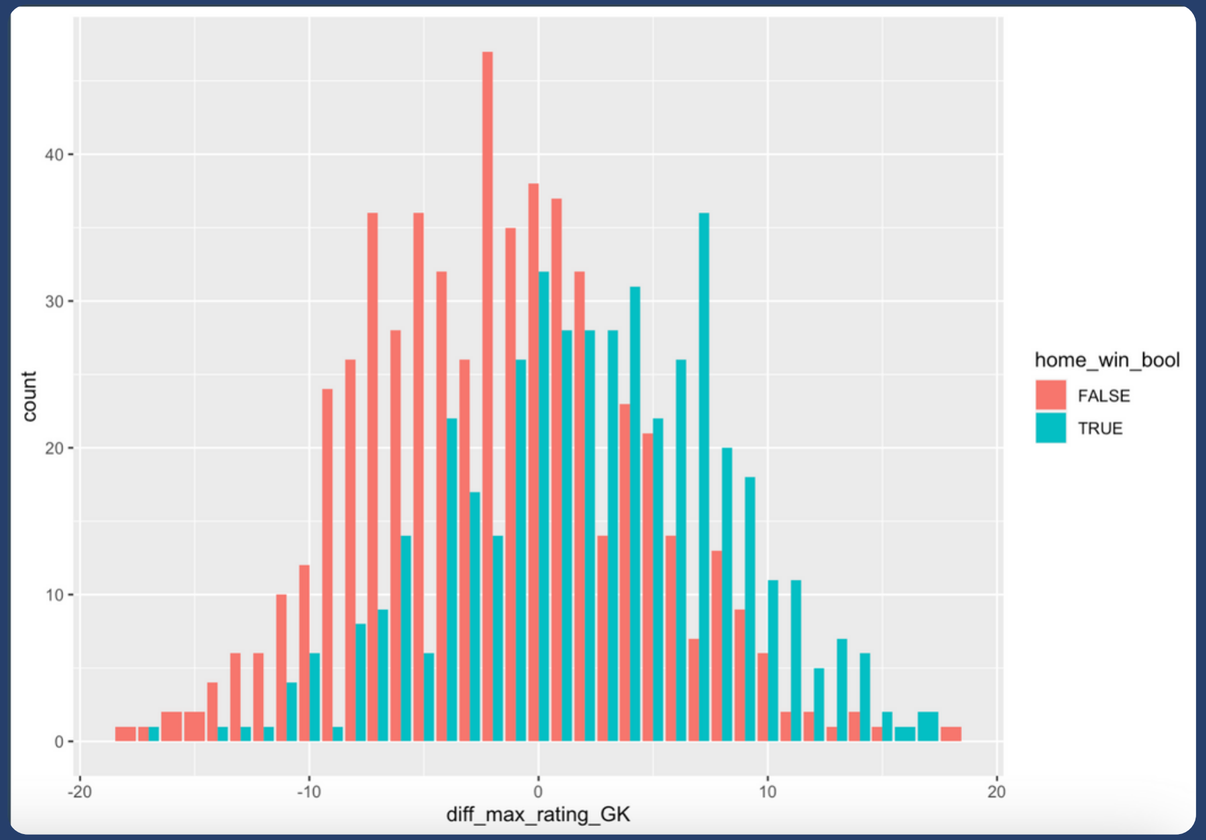
- merging by club names
- home and away teams
- home and away win and losses

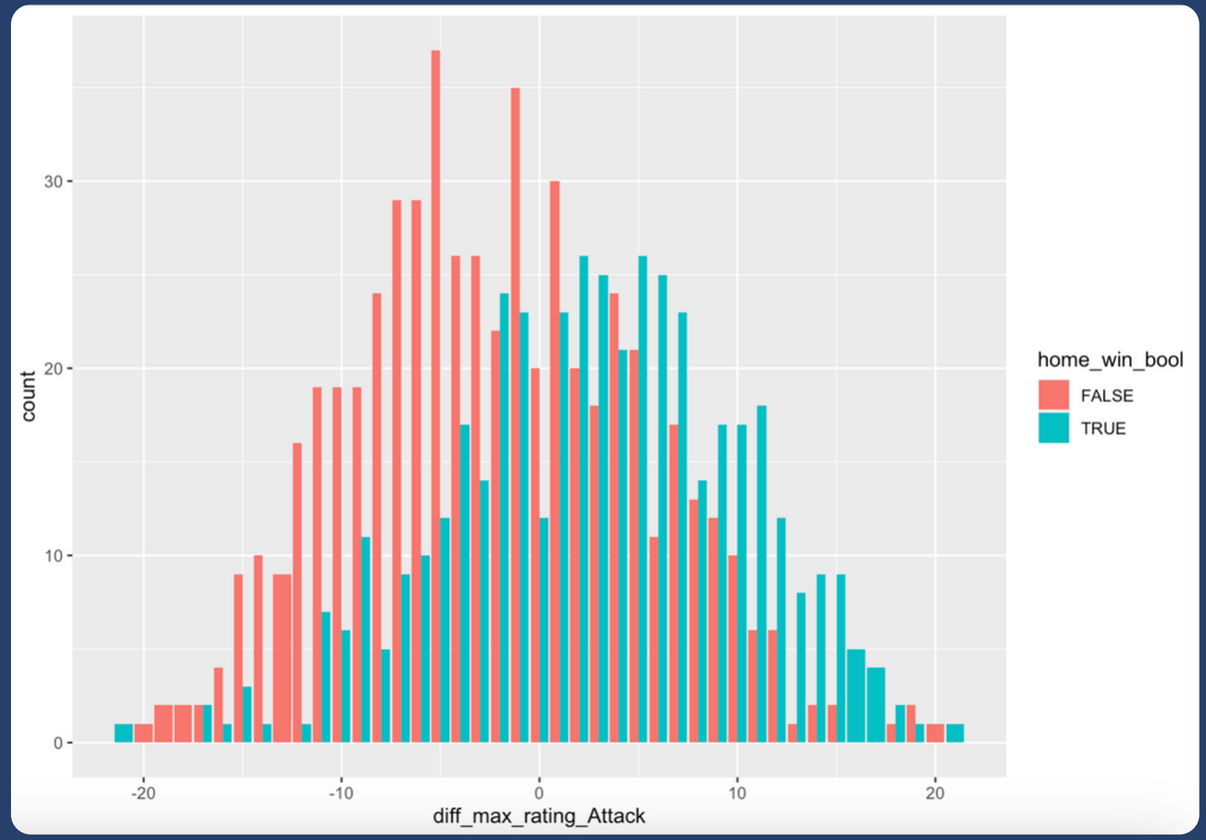
=


## FINAL DATASET

Final dataset used for our logistic regression model.

# DATA SUMMARIES:



# PROBLEMS WE HOPE TO SOLVE

**Project Objective:**

- Assess the reliability of using FIFA ratings in predicting Premier League match outcomes.
- The significance lies in aiding sports bettors, particularly in moneyline betting scenarios.

**Moneyline Betting Overview:**

- Moneyline bets focus on whether a team will win or not.
- Positive values indicate underdogs; e.g., +145 means a $100 bet yields $145 if the team wins.
- Negative values imply favorites; e.g., -300 requires a $300 bet for a $100 payout if they win.

**Preseason Betting Insights:**

- Predictions aid preseason bets on Premier League winners or relegations.
- Wins contribute to league standings, influencing season outcomes.

**Decision Support for Bettors:**

- Developing predictive models aids bettors in decision-making.
- Predicting match outcomes, factoring in team matchups, assists in maximizing winnings or minimizing losses in sports betting.

# ANALYTICAL TASK: Prediction

## Model 1

```
glm(formula = home_win_1_0 ~ ., family = binomial(link = "logit"),
    data = premier_league_results_merged_fifa_rating_summary_stats[,
        c("home_win_1_0", colnames_holder_filt)])

Coefficients: (2 not defined because of singularities)
                                          Estimate Std. Error z value
(Intercept)                             -0.230548   0.200547  -1.150
diff_max_rating_Defense                  0.006635   0.108690   0.061
diff_max_rating_GK                       0.062816   0.035631   1.763
diff_max_rating_Midfield                -0.047616   0.038311  -1.243
diff_max_rating_Attack                   0.092200   0.081020   1.138
diff_median_rating_Defense              -0.036475   0.110795  -0.329
diff_median_rating_Midfield              0.091401   0.043416   2.105
diff_median_rating_Attack               -0.106982   0.080096  -1.336
diff_var_rating_Defense                 -0.003235   0.010343  -0.313
diff_var_rating_Midfield                 0.001344   0.007897   0.170
diff_var_rating_Attack                  -0.045337   0.020061  -2.260
diff_sum_work_rate_Defense               0.001802   0.076106   0.024
diff_sum_work_rate_GK                    0.057999   0.023069   2.514
diff_sum_work_rate_Midfield              0.021932   0.065295   0.336
diff_sum_work_rate_Attack                0.077553   0.081822   0.948
diff_median_rating_home_attack_away_defense  0.010364   0.032316   0.321
diff_median_rating_home_defense_away_attack        NA         NA      NA
diff_max_rating_home_attack_away_gk      0.032468   0.042087   0.771
diff_max_rating_home_gk_away_attack            NA         NA      NA
                                        Pr(>|z|)
(Intercept)                                0.250
diff_max_rating_Defense                    0.951
diff_max_rating_GK                         0.078 .
diff_max_rating_Midfield                   0.214
diff_max_rating_Attack                     0.255
diff_median_rating_Defense                 0.742
diff_median_rating_Midfield                0.035 *
diff_median_rating_Attack                  0.182
diff_var_rating_Defense                    0.754
diff_var_rating_Midfield                   0.865
diff_var_rating_Attack                     0.024 *
diff_sum_work_rate_Defense                 0.981
diff_sum_work_rate_GK                      0.012 *
diff_sum_work_rate_Midfield                0.737
diff_sum_work_rate_Attack                  0.343
diff_median_rating_home_attack_away_defense  0.748
diff_median_rating_home_defense_away_attack      NA
diff_max_rating_home_attack_away_gk        0.440
diff_max_rating_home_gk_away_attack              NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1376.5  on 1001  degrees of freedom
Residual deviance: 1220.7  on  985  degrees of freedom
AIC: 1254.7

Number of Fisher Scoring iterations: 4
```

## Model 1 Confusion Matrix

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 110  33
         1  42  66

               Accuracy : 0.7012
                 95% CI : (0.6404, 0.7571)
    No Information Rate : 0.6056
    P-Value [Acc > NIR] : 0.001025

                  Kappa : 0.3843

 Mcnemar's Test P-Value : 0.355611

            Sensitivity : 0.6667
            Specificity : 0.7237
         Pos Pred Value : 0.6111
         Neg Pred Value : 0.7692
             Prevalence : 0.3944
         Detection Rate : 0.2629
   Detection Prevalence : 0.4303
      Balanced Accuracy : 0.6952

       'Positive' Class : 1
```

# ANALYTICAL TASK: Prediction

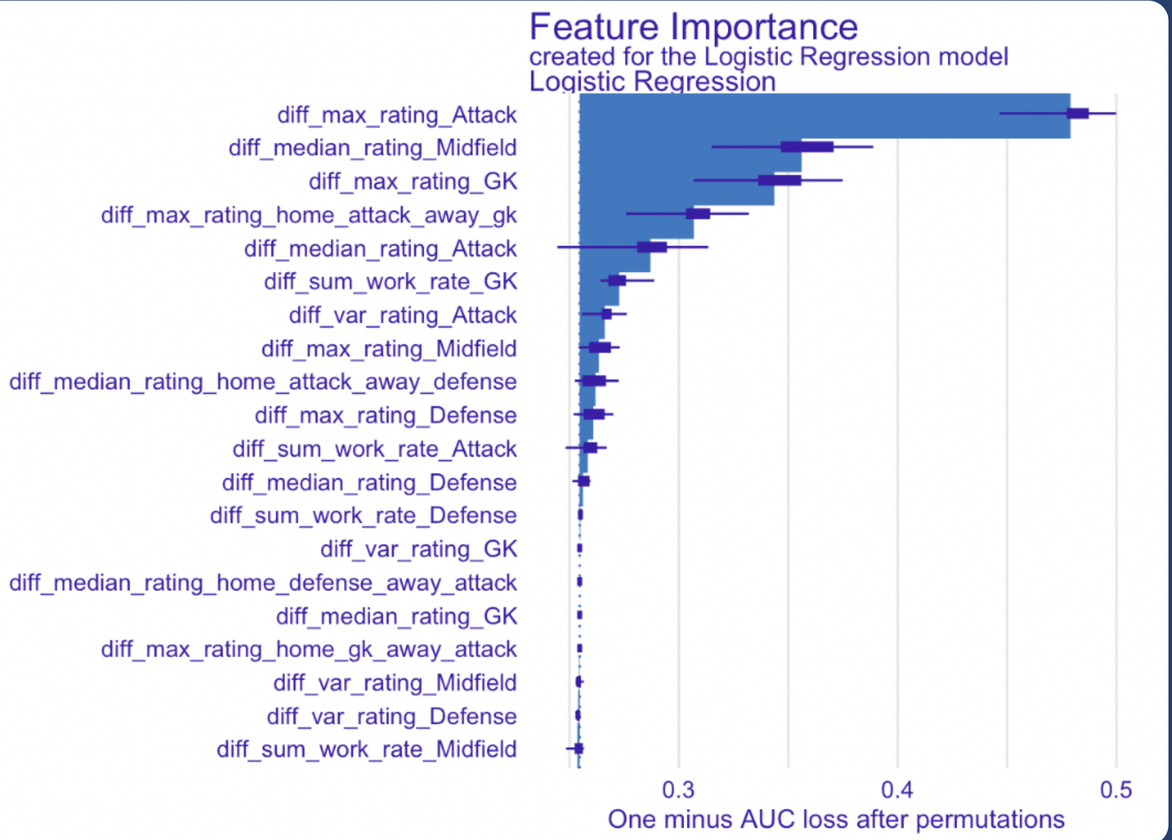
## Model 2

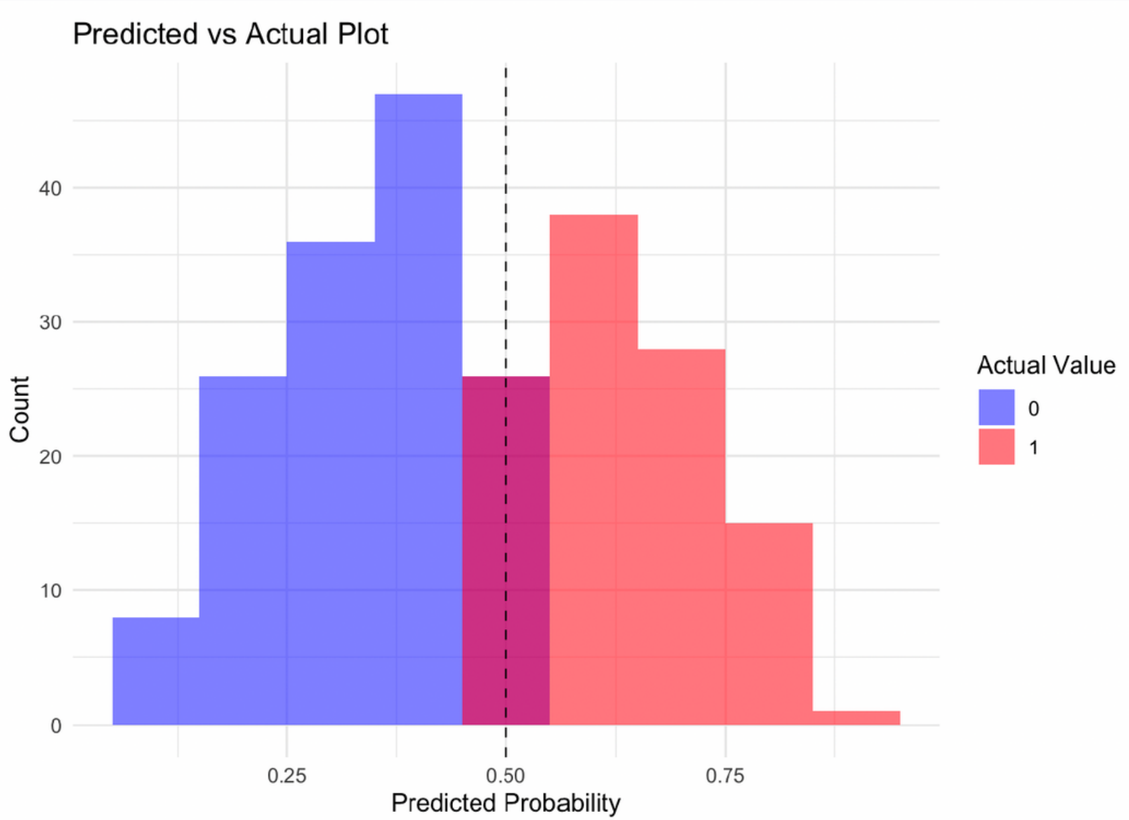
```
glm(formula = home_win_1_0 ~ diff_median_rating_home_attack_away_defense +
    diff_median_rating_home_defense_away_attack + diff_max_rating_home_attack_away_gk +
    diff_max_rating_home_gk_away_attack + diff_median_rating_Midfield,
    family = binomial(link = "logit"), data = premier_league_results_merged_fifa_rating_summary_stats[,
        c("home_win_1_0", colnames_holder_filt)])

Coefficients:
                                          Estimate Std. Error z value
(Intercept)                               -0.22333    0.19866  -1.124
diff_median_rating_home_attack_away_defense -0.02088    0.02464  -0.847
diff_median_rating_home_defense_away_attack -0.03094    0.02453  -1.261
diff_max_rating_home_attack_away_gk          0.07511    0.02629   2.857
diff_max_rating_home_gk_away_attack          0.04274    0.02630   1.625
diff_median_rating_Midfield                  0.05405    0.02315   2.335
                                          Pr(>|z|)
(Intercept)                                0.26094
diff_median_rating_home_attack_away_defense  0.39673
diff_median_rating_home_defense_away_attack  0.20719
diff_max_rating_home_attack_away_gk          0.00427 **
diff_max_rating_home_gk_away_attack          0.10417
diff_median_rating_Midfield                  0.01954 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1376.5  on 1001  degrees of freedom
Residual deviance: 1241.7  on  996  degrees of freedom
AIC: 1253.7

Number of Fisher Scoring iterations: 4
```

## Model 2 Confusion Matrix

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 108  36
         1  44  63

               Accuracy : 0.6813
                 95% CI : (0.6197, 0.7385)
    No Information Rate : 0.6056
    P-Value [Acc > NIR] : 0.007858

                  Kappa : 0.3421

 Mcnemar's Test P-Value : 0.433848

            Sensitivity : 0.6364
            Specificity : 0.7105
         Pos Pred Value : 0.5888
         Neg Pred Value : 0.7500
             Prevalence : 0.3944
         Detection Rate : 0.2510
   Detection Prevalence : 0.4263
      Balanced Accuracy : 0.6734

       'Positive' Class : 1
```

# DATA VISUALIZATIONS



# PREDICTIVE ACCURACY

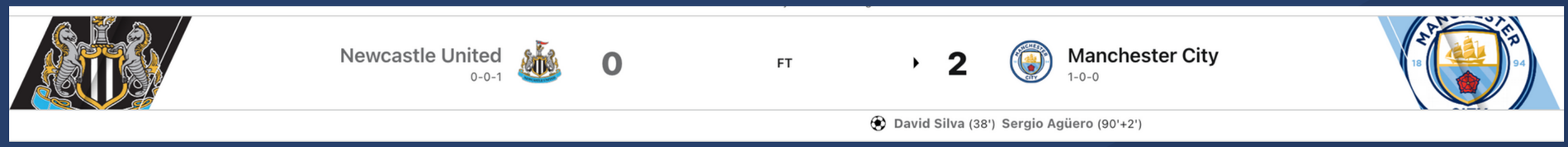
| Cutoff of 0.5 | Model 1 | Model 2 |
|---|---|---|
| Accuracy | **0.7012** | 0.6813 |
| Sensitivity (True Positive Rate) | **0.6667** | 0.6364 |
| Specificity (True Negative Rate): | **0.7237** | 0.7105 |
| Positive Predictive Value (Precision): | **0.6111** | 0.5888 |
| Negative Predictive Value: | 0.7500 | **0.7692** |
| Detection Rate: | **0.2629** | 0.2510 |
| Balanced Accuracy: | **0.6952** | 0.6734 |

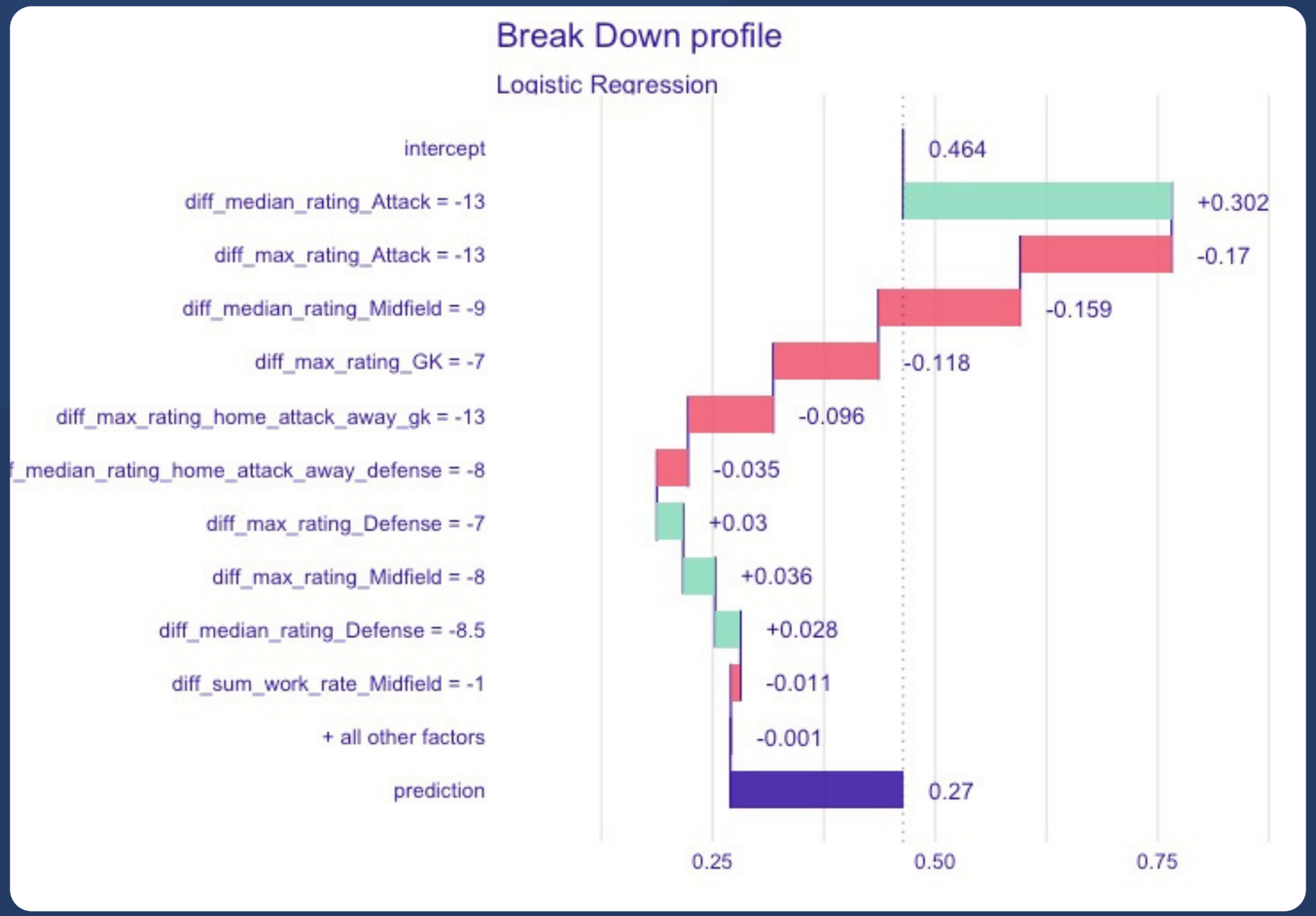
**Model 1 is best**

# PREDICTIVE ACCURACY

| Model | Cutoff of 0.5 | Cutoff 0.6 |
|---|---|---|
| Accuracy | 0.6972 | **0.7211** |
| Sensitivity (True Positive Rate) | **0.6465** | 0.4646 |
| Specificity (True Negative Rate): | 0.7303 | **0.8881** |
| Positive Predictive Value (Precision): | 0.6095 | **0.7302** |
| Negative Predictive Value: | 0.7603 | **0.7181** |
| Detection Rate: | **0.2550** | 0.1832 |
| Balanced Accuracy: | **0.6884** | 0.6764 |

# IMPLIMENTATION EXAMPLE




# FINDINGS

**Performance Metrics in Betting:**

- **Specificity** is crucial, highlighting false positives (incorrect predictions of a team's win).
  - Higher specificity, especially with a 0.6 cutoff, is desirable for minimizing incorrect win predictions.
- **Positive Predictive Value (PPV)** measures accuracy in predicting home team wins.
  - A 0.6 cutoff in Model 1 enhances PPV, making it more desirable for maximizing win predictions.
- **Bet Type Consideration:**
  - Different bettor objective: maximize winnings or minimize losses.
- **For maximizing wins, focus on PPV**, favoring Model 1 with a 0.6 cutoff for higher correct win predictions.
- **For minimizing losses, prioritize specificity**, once again favoring Model 1 with a 0.6 cutoff for fewer incorrect win predictions.

# CONCLUSION

**Model Insights for Betting:**

- Premier League betting using FIFA data emphasizes the significance of differences in attack, midfield, and goalkeeper ratings.
- Matchups (ex : home attack vs. away defense) not as important as expecetd
- Work rate variations seem to be less influential than expected
- FIFA ratings serve as a useful tool for creative insights into team capabilities and predicting game outcomes