KOGOD SCHOOL of BUSINESS
AMERICAN UNIVERSITY • WASHINGTON, DC

ITEC 621 Predictive Analytics Project

Project Name: Predicting Total Runs in Major League Baseball (MLB) Games

Class Section: 002

Team Number: 1

Team Members:

Eric Baran

Edwin Bwambale

Wasif Talukder

Nabila Raisa

1. Business Question and Case

1.1 Business Question

We seek to answer whether a Major League Baseball (MLB) game's total score can be predicted using historical team performance and start of game data. A model predicting whether a game's total score will be above or below a certain score will be useful for sports betting context.

1.2 Business Case

In 2018, US Supreme Court overturned the Professional and Amateur Sports Protection Act [1]. Five years since, 37 states plus D.C. has wagering activities in amateur or professional sports [2]. In 2023, the total amount of money bet on sports reached $119.8 billion [3]. Therefore, a model predicting "totals" – a game's combined score – can be used by sports books, sports betters and sports fans for setting wagers, taking bets or finding entertaining high scoring games.

2. Analytics Question

The analytics question is whether a future MLB game's total score will be "over" or "under" the average baseball game score; currently, this is around 8 or 9 total runs. Inputs for prediction will include historical team performance and game start conditions. Our model prioritizes predicting correctly over interpretability. Model performance measurements include accuracy and ROC-AUC, and we will maximize accuracy.

2.1 Outcome Variable of Interest

The outcome variable of interest is whether the total runs scored in a game is over or under 8.5 total runs. This is a binary variable of "1" for above and "0" for under 8.5.

2.2 Main Predictors

The model considers factors likely to affect total runs scored. Generally, the factors cover historical performance for the two teams in a matchup and conditions about start of game.

Firstly, Predictors for historical performance are measured based on the average performance from the most recent last ten games, and these are quantitative variables, equal to or greater than 0. Separate predictors are included to measure both the home and away team. Specific predictors are runs scored, runs allowed, hits, errors, players left on base, times at bat, doubles, triples, and homeruns. Additional variables cover the teams' win percentage, games behind division leader, and last team matchup's totals.

Secondly, specific predictors for game start include nighttime game (binary), doubleheader (categorical), precipitation (binary), sunny (binary), temperature (quantitative), wind speed (quantitative, ≥ 0), wind direction (categorical), and park (binary for Coors field in Denver). These variables may contribute to the total score in a game. For example, balls travel farther when temperature and humidity climb, which may result in more runs scored in a game [4, 5].

3. Data Set Description

Four large MLB datasets are sourced from “baseball.computer” and filtered for the 2021, 2022 and 2023 regular seasons. There are 30 teams in the MLB and in the original dataset. Over the regular season, 162 games are scheduled, and 7,289 total games are in the original dataset (one less than 7,290 possible). These four datasets are joined and structured for one row per game.

The compiled dataset calculates the within season historical team performance based on the last teams’ matchup or each team’s last 10 games. Therefore, within each season, the first team matchup and the first 10 games played by each team won’t have data and are dropped from the final dataset. The final dataset contains 5,947 games.

4. Descriptive Analytics

4.1 Descriptive statistics of key variables

Average temperature (Appendix 7) is approximately 74 F (degree Fahrenheit) with a standard deviation of 9.55 i.e. about 95% of games are played between the temperatures of 55F-93F (between 2SD). Average windspeed for MLB regular season games is 6.75mph, which is considered a very gentle breeze according to weather experts. Other interesting findings are hometeam and away time average win percentages are at around 50%, making our analysis agnostic of hometeam advantages.

4.2 Distribution of key variables

Two of the most important game start condition quantitative variables are:

a) Temperature (temperature_fahrenheit) ranges from 34F to 109F. This variable is left skewed as indicated with the long left tail for this variable’s histogram and deviation from theoretical normality in the QQ plot (Appendix 2).

b) Wind speed (wind_speed_mph) ranges from 0 mph to 36 mph. It is bimodal at 0 mph and around 7 mph and is right skewed as indicated with the long right tail for this variable’s histogram and deviation from theoretical normality in the QQ plot (Appendix 3).

4.3 Correlation and co-variation analysis

Chi-squared tests were performed for two most important categorical game start condition variables:

a) “Coors Field”: significant correlation was discovered with a p-value <0.001 (Appendix 4)

b) For the first and second doubleheader games (Appendix 5 and 6), significant correlation also was discovered where, on average, double header games were less likely to have totals above 8.5 (p-value <0.05 for both first and second double header game.

For the quantitative team performance variables, a correlation analysis was performed. (Appendix 8). Moderate positive correlations were found between bats and hits, home runs and runs scores, hits and doubles, and wins and runs scored. Moderate negative correlation was found between games behind and win percentage and runs allowed and win percentage. These correlations make intuitive sense.

4.4 Data preprocessing and transformations

The outcome variable totals_above_average was a transformation from quantitative to binary. In the full dataset of 5,947 games, 3,026 fell below 8.5 runs and 2,921 observations fell above 8.5 runs, which makes this a balanced dataset. Most current totals bets on game day before game start are split at or around 8.5, which makes this a useful split for betting (Appendix 1).

5. Modeling Methods and Model Specifications

5.1. Initial Model Specification

The initial model specification includes all 34 predictor variables without transformation included in the dataset description and the target variable for whether a game will exceed 8.5 runs. Additionally, the model includes a few interaction terms. For game start conditions, the model includes interactions between wind direction and wind speed and temperature and time of game. For historical team performance, the model includes interactions between runs scored, runs allowed and home runs for the home and away teams.

5.2. Initial OLS or Logit Model Results

A logit model (Appendix 9) was fit using these predictors found multiple statistically significant variables at or below the 0.1 level. Significant predictors include doubleheader, precipitation, wind direction, temperature, wind speed, the away team's prior hits, the away teams at bats, the home team's games behind, the prior teams' matchup total, and certain playing fields. Additionally, interactions between wind direction and wind speed were found significant below the 0.05 level. Some ballparks were also found more significant, such as Denver. This initial model reduces the null model's deviance (or 2LL) at 9,593.1 to residual deviance at 9,230.6, which is little reduction in error.

5.3 Assumption Tests

Since this is a classification model, we will test for multicollinearity by inspecting the condition index (C.I.) and serial correlation using a Durbin-Watson (DW) test.

Multicollinearity tests (Appendix 10): The largest C.I. is found at 313. Since this is above 50, multicollinearity exists and should be addressed. We also consulted the Variance Inflation Factors (V.I.F) and besides interaction terms, VIFs were found less than 10. This suggests that predictors between themselves do not have a high linear association (Appendix 10), and we may not need to specifically address multicollinearity. However, we consider shrinkage and dimensionality reduction to address some multicollinearity as well as to remove variables without significant influence in the model.

Serial correlation test (Appendix 11): The data was grouped by home team, sorted by date of game, and a DW test was performed. The resulting DW test statistic is at 2.0012. Since this near 2, no serial correlation is detected.

5.4. Model Candidates and Rationale

The three model candidates selected were generalized logistic regression, random forest, and boosted trees. We considered these three models to test both a parametric approach and a non-parametric approach for predicting the outcome. Within generalized logistic regression, we used Lasso to reduce the number of features to address multicollinearity and dimensionality. In addition, Lasso will drop predictors by shrinking their coefficients to 0 that do not have significant predictive power. We chose random forest and boosted trees due to their training approach, such as bootstrapped samples and fitting multiple models, that often yield higher accuracy, compared to a traditional decision tree. Appendix 12 includes the three models and the two specifications that are addressed in the next section.

5.5. Model Specification Candidates and Rationale

The model specifications used were based on business knowledge and dimensionality reduction. For business knowledge, all predictors were used that are expected to result impact game totals. For example, more times a team is at bat and more hits are likely to result in more runs. For dimensionality reduction, we use principal component analysis.

5.6. Cross-Validation Testing and Final Model Selection

a) Generalized Logistic using LASSO to shrink and drop variables: accuracy at 53% is not a noticeable improvement over a coin toss.

b) PCR to reduce dimensions: accuracy is bad at 45% i.e. less than 50% chance level.

The first Random Forest through business knowledge based feature selection had an accuracy of 0.53096 or 53.10%. The tuning parameter mt was set to 2. The second Random Forest was done with statistically selected predictors and tuning parameter mt=6. Accuracy improved by about half a percent to 0.53567 or 53.57%

The first Boosted Tree with business knowledge based variable selection had an accuracy of 57%. The 10 fold cross-validation boosting used 116 stumps (single-split trees) and a learning rate of 0.05. A second Boosted Tree with statistically selected features did not outperform the first Boosted Tree model (55%) but did perform better than other model-specifications tested.

Finally, we can say that based on Cross-Validated Accuracy Rates (Appendix 12), the best model was Boosted Tree with business knowledge based variable selection with accuracy of 57%. It has a 7% better performance than chance level accuracy rate.

6. Analysis of Results

Our best model's confusion matrix shows that our model classifies the total score prediction as under more often than over (3,619 vs 2,328) even though the actual data has a balanced split – 3,026 over and 2,921 under. (Appendix 13). This results in an overall accuracy at 57%. This is the metric we are maximizing. Additionally, the model has a sensitivity of 46% and a specificity of 67% at a 50% threshold (Appendix 14).

From the Variance Importance Plot (Appendix 13), we find the most influential variable to be temperature. Temperature was also found to be significant in the other models, such as Random

Forest. (Appendix 15). Intuitively, this makes sense since warmer air is less dense, which carries a ball further. Similarly, Coors Field in Denver was also found to be very important. Coors field in the highest altitude stadium, and because of its higher altitude, the air is also less dense, which carries a ball further.

Additionally, many of the home team variables were found in the feature importance plot, whereas five away team variables have no influence. This suggests the historical home team is more influential in total runs scored compared to the visiting away team. This could make sense under a home team advantage rationale.

A sunny sky, a nighttime game and win across the field were not found important in this model, which is in contrast to the Random Forest model (Appendix 15).

7. Conclusions and Lessons Learned

7.1. Conclusions from the Analysis

The findings from our analysis suggest predicting the total score of Major League Baseball (MLB) games using historical team performance and start-of-game data does better than even chance. Specifically, our model outperforms by 7% compared to random guessing. The most influential variables using the Boosted Tree model with all predictors were found to be start of game temperature, the home teams' latest win percentage, games played at Coors field, average 10 games' runs allowed by the home team, and average 10 games' hits by the away team.

Although, setting wagers or betting on margins slightly above 50% is not feasible in practice. Successfully winning bets on 57 games out of 100 does not result significant gain. However, the Boosted Tree model provides some interpretation about feature importance, which can be leveraged as new business knowledge in future models. We also encourage use of our model for predicting "under" bets for regular season MLB games as it has a 67% specificity.

7.2. Project Issues, Challenges and Lessons Learned

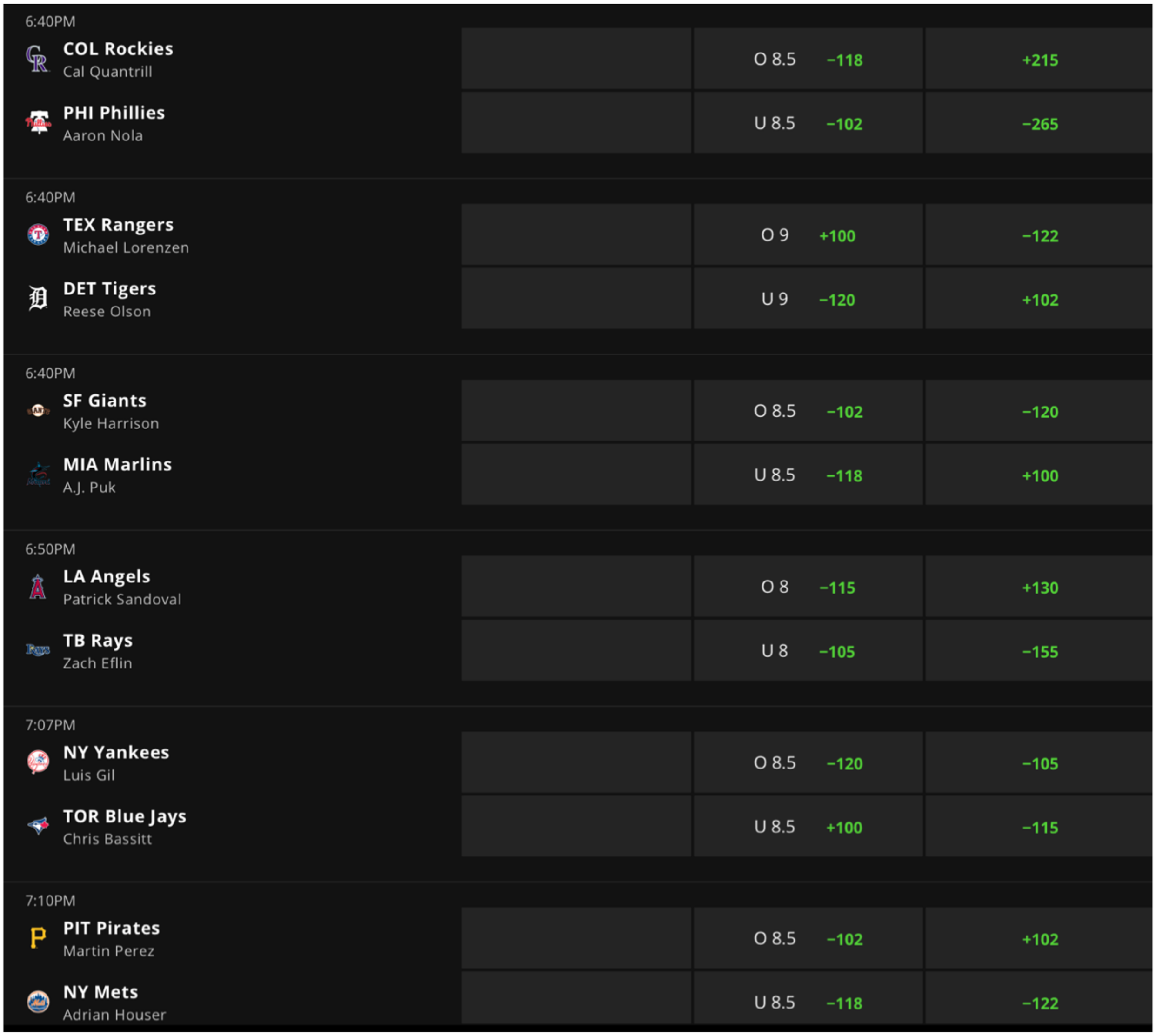
There were three issues, challenges or lessons learned in this project. First, we shifted to framing the problem as a classification task (predicting "over" or "under") rather than a regression task (predicting actual total scores). This was a more appropriate approach given the problem statement and intended use case. Second, we should consider additional factors such as individual player performance and team lineup, which are likely to influence game scores. Identifying the most relevant variables from the available data proved is challenging, and we will have to continue exploring predictors that capture the nuances of for total runs.

Leveraging domain expertise in baseball and sabermetrics is essential for refining the model and selecting meaningful predictors and predictive models. This highlights the importance of collaboration between data analysts and domain experts to ensure the relevance and accuracy of the predictions.

Appendix

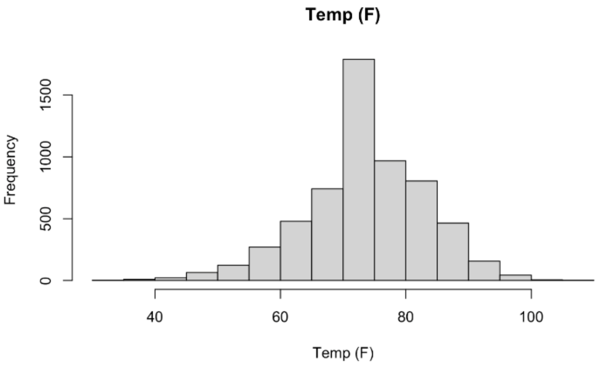
1. Data over-under split & Draft King’s Baseball Wagers prior to Game Start (4/13/24)

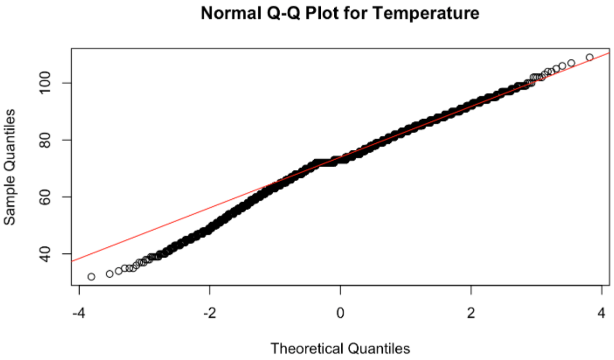
Total score over-under 8.5		
Under	Over	Sum
3026	2921	5,947



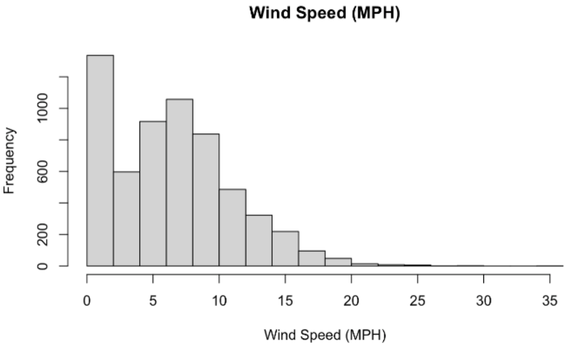
6:40PM	COL Rockies	Cal Quantrill		O 8.5	−118	+215
	PHI Phillies	Aaron Nola		U 8.5	−102	−265
6:40PM	TEX Rangers	Michael Lorenzen		O 9	+100	−122
	DET Tigers	Reese Olson		U 9	−120	+102
6:40PM	SF Giants	Kyle Harrison		O 8.5	−102	−120
	MIA Marlins	A.J. Puk		U 8.5	−118	+100
6:50PM	LA Angels	Patrick Sandoval		O 8	−115	+130
	TB Rays	Zach Eflin		U 8	−105	−155
7:07PM	NY Yankees	Luis Gil		O 8.5	−120	−105
	TOR Blue Jays	Chris Bassitt		U 8.5	+100	−115
7:10PM	PIT Pirates	Martin Perez		O 8.5	−102	+102
	NY Mets	Adrian Houser		U 8.5	−118	−122

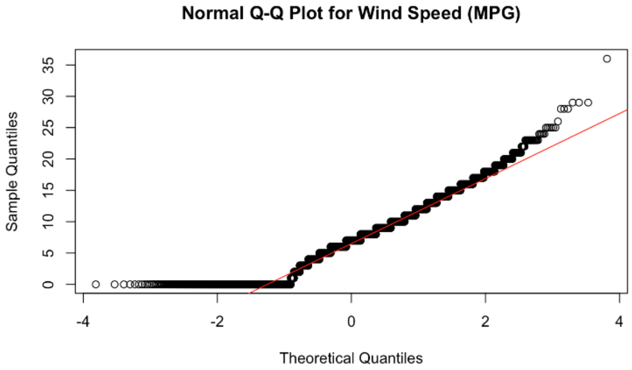
2. Histogram and QQ Plot for Temperature





3. Histogram and QQ Plot for Wind Speed





4. Contingency Table and Chi Squared Test for Coors Field and Game Total’s above 8.5.

```
     0    1
0 2971 2787
1   55  134

Pearson's Chi-squared test with Yates' continuity correction

data:  data
X-squared = 36.164, df = 1, p-value = 0.000000001814
```

5. Confusion Matrix and chi-squared test for first double header game and total’s above 8.5.

```
       0    1
  0 2939 2866
  1   87   55

Pearson's Chi-squared test with Yates' continuity correction

data:  data
X-squared = 5.8589, df = 1, p-value = 0.0155
```

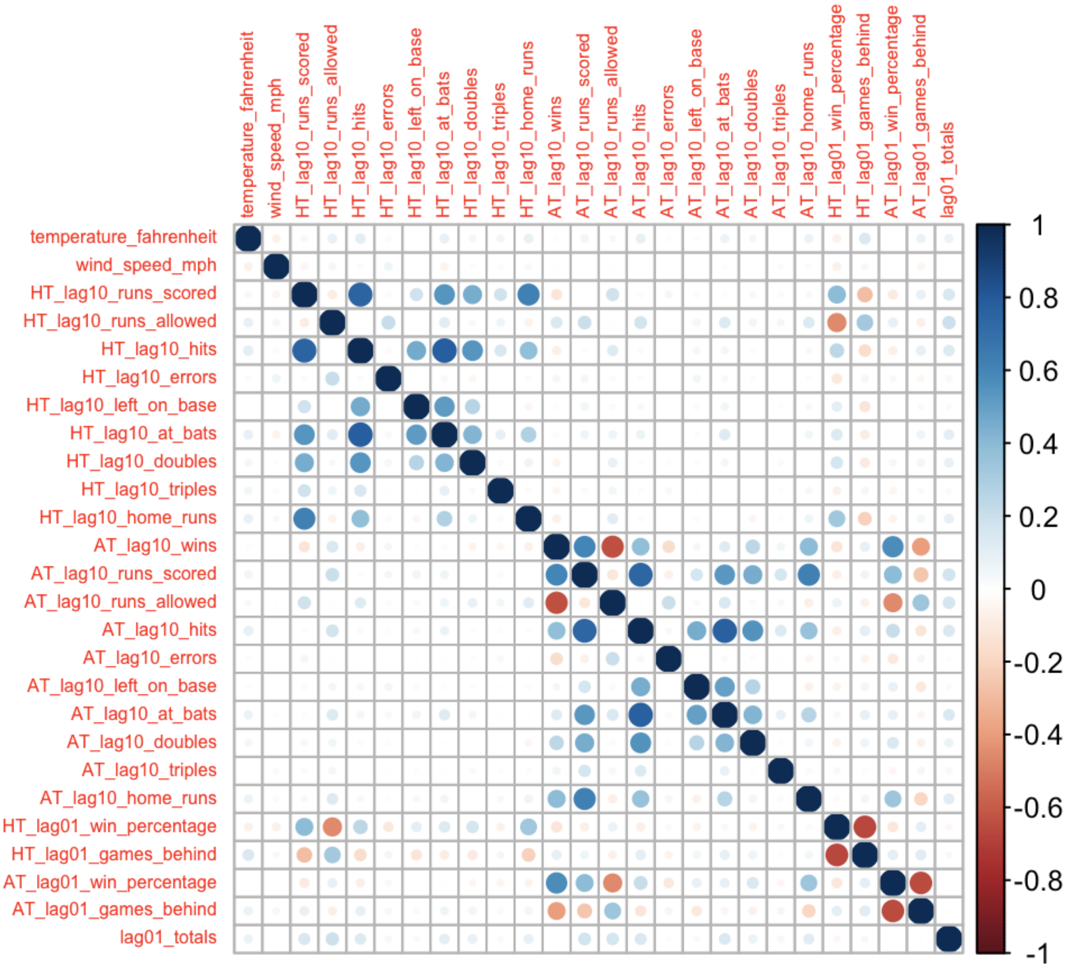
6. Confusion Matrix and chi-squared test for second double header game and total’s above 8.5.

```
       0    1
  0 2936 2862
  1   90   59

Pearson's Chi-squared test with Yates' continuity correction

data:  data
X-squared = 5.1582, df = 1, p-value = 0.02314
```

7. Mean & Standard deviations

Variable	Mean	SD
temperature_fahrenheit	74.11266185	9.55
wind_speed_mph	6.749285354	4.88
HT_lag10_runs_scored	4.482713973	1.08
HT_lag10_runs_allowed	4.495426265	1.12
HT_lag10_hits	8.259895746	1.16
HT_lag10_errors	0.513435346	0.25
HT_lag10_left_on_base	6.6267698	0.87
HT_lag10_at_bats	33.71989238	1.33
HT_lag10_doubles	1.657087607	0.45
HT_lag10_triples	0.136690768	0.13
HT_lag10_home_runs	1.180191693	0.42
AT_lag10_wins	0.505179082	0.18
AT_lag10_runs_scored	4.495913906	1.07
AT_lag10_runs_allowed	4.477080881	1.13
AT_lag10_hits	8.245888683	1.15
AT_lag10_errors	0.510038675	0.25
AT_lag10_left_on_base	6.603480747	0.88
AT_lag10_at_bats	33.54570372	1.31
AT_lag10_doubles	1.652900622	0.45
AT_lag10_triples	0.140289221	0.13
AT_lag10_home_runs	1.172961157	0.42
HT_lag01_win_percentage	0.500122249	0.10
HT_lag01_games_behind	9.189086935	9.68
AT_lag01_win_percentage	0.500184283	0.10
AT_lag01_games_behind	9.213300824	9.64
lag01_totals	8.934252564	4.48

8. Correlation matrix

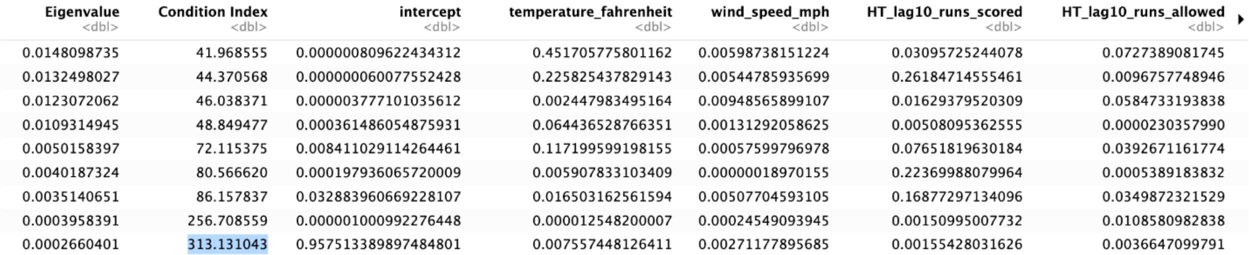


9. Output From Full Logistic Regression Model

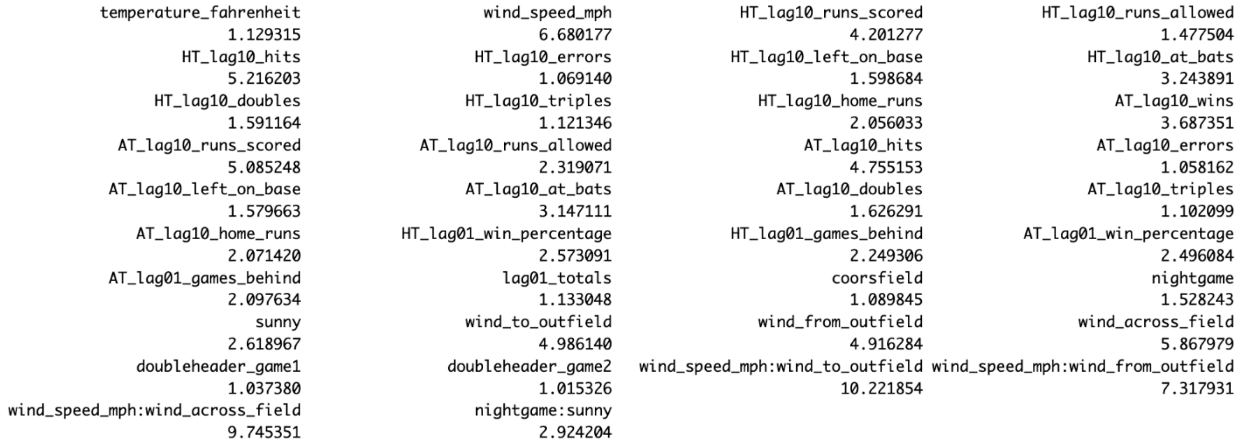
```
Coefficients:
                                             Estimate Std. Error z value
Pr(>|z|)
(Intercept)                        -0.8124561  1.2145205  -0.669          0.50353
temperature_fahrenheit              0.0204974  0.0029713   6.898 0.0000000000526 ***
wind_speed_mph                     -0.0006084  0.0139308  -0.044          0.96516
HT_lag10_runs_scored                0.0086136  0.0496332   0.174          0.86222
HT_lag10_runs_allowed               0.0837074  0.0286566   2.921          0.00349 **
HT_lag10_hits                      -0.0117912  0.0515725  -0.229          0.81915
HT_lag10_errors                    -0.1008417  0.1096620  -0.920          0.35780
HT_lag10_left_on_base               0.0138713  0.0382649   0.363          0.71697
HT_lag10_at_bats                   -0.0178451  0.0356626  -0.500          0.61680
HT_lag10_doubles                    0.1066448  0.0733467   1.454          0.14595
HT_lag10_triples                   -0.2021050  0.2224045  -0.909          0.36349
HT_lag10_home_runs                  0.1374371  0.0900935   1.525          0.12714
AT_lag10_wins                      -0.1041019  0.2838071  -0.367          0.71376
AT_lag10_runs_scored               -0.0365872  0.0552362  -0.662          0.50773
AT_lag10_runs_allowed               0.0039472  0.0352753   0.112          0.91091
AT_lag10_hits                       0.0943128  0.0500590   1.884          0.05956 .
AT_lag10_errors                    -0.1443155  0.1101540  -1.310          0.19015
AT_lag10_left_on_base               0.0384819  0.0375687   1.024          0.30569
AT_lag10_at_bats                   -0.0419078  0.0356491  -1.176          0.23977
```

```
AT_lag10_doubles                  -0.0226500  0.0745621  -0.304          0.76130    
AT_lag10_triples                   0.1309865  0.2152644   0.608          0.54286    
AT_lag10_home_runs                 0.0503727  0.0898157   0.561          0.57490    
HT_lag01_win_percentage           -0.0380077  0.4388767  -0.087          0.93099    
HT_lag01_games_behind              0.0038409  0.0040931   0.938          0.34805    
AT_lag01_win_percentage           -0.3379217  0.4314389  -0.783          0.43348    
AT_lag01_games_behind             -0.0052445  0.0039526  -1.327          0.18457    
lag01_totals                       0.0007059  0.0062617   0.113          0.91024    
coorsfield                         0.8772313  0.1723359   5.090 0.00000035760840 ***
nightgame                         -0.0398726  0.0662264  -0.602          0.54713    
sunny                             -0.1192818  0.0913703  -1.305          0.19173    
wind_to_outfield                   0.0064750  0.1263474   0.051          0.95913    
wind_from_outfield                -0.0143909  0.1558876  -0.092          0.92645    
wind_across_field                 -0.3706559  0.1427823  -2.596          0.00943 ** 
doubleheader_game1                -0.4500939  0.1795781  -2.506          0.01220 *  
doubleheader_game2                -0.3601303  0.1730509  -2.081          0.03743 *  
wind_speed_mph:wind_to_outfield    0.0124629  0.0178656   0.698          0.48543    
wind_speed_mph:wind_from_outfield -0.0001255  0.0214640  -0.006          0.99533    
wind_speed_mph:wind_across_field   0.0434436  0.0198039   2.194          0.02826 *  
nightgame:sunny                    0.1101919  0.1152714   0.956          0.33911    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8242.4  on 5946  degrees of freedom
Residual deviance: 8080.8  on 5908  degrees of freedom
AIC: 8158.8

Number of Fisher Scoring iterations: 4
```

10. Condition Index > 50 in Full Model and VIF



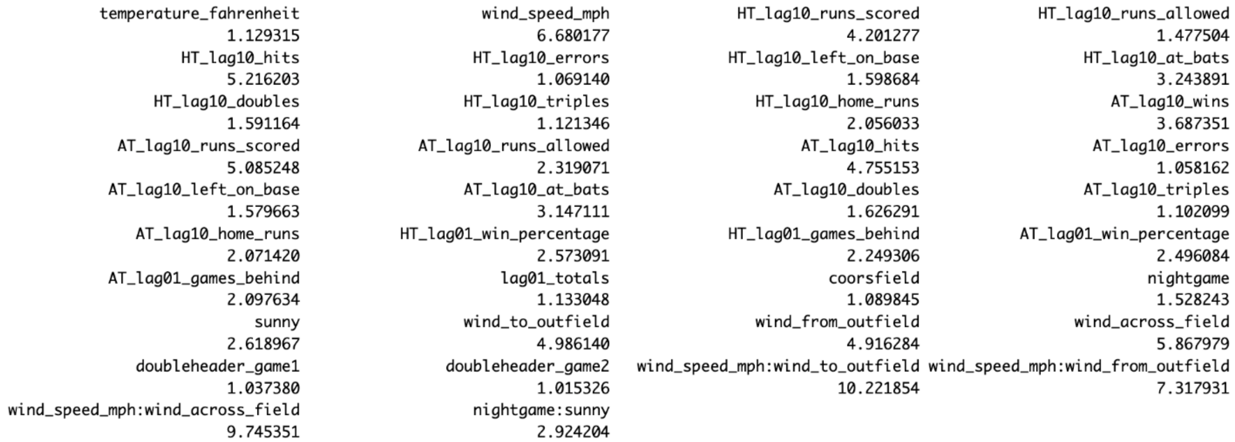
Eigenvalue <dbl>	Condition Index <dbl>	intercept <dbl>	temperature_fahrenheit <dbl>	wind_speed_mph <dbl>	HT_lag10_runs_scored <dbl>	HT_lag10_runs_allowed <dbl>
0.0148098735	41.968555	0.000000809622434312	0.451705775801162	0.00598738151224	0.03095725244078	0.0727389081745
0.0132498027	44.370568	0.000000060077552428	0.225825437829143	0.00544785935699	0.26184714555461	0.0096757748946
0.0123072062	46.038371	0.000003777101035612	0.002447983495164	0.00948565899107	0.01629379520309	0.0584733193838
0.0109314945	48.849477	0.000361486054875931	0.064436528766351	0.00131292058625	0.00508095362555	0.0000230357990
0.0050158397	72.115375	0.008411029114264461	0.117199599198155	0.00057599796978	0.07651819630184	0.0392671161774
0.0040187324	80.566620	0.000197936065720009	0.005907833103409	0.00000018970155	0.22369988079964	0.0005389183832
0.0035140651	86.157837	0.032883960669228107	0.016503162561594	0.00507704593105	0.16877297134096	0.0349872321529
0.0003958391	256.708559	0.000001000992276448	0.000012548200007	0.00024549093945	0.00150995007732	0.0108580982838
0.0002660401	313.131043	0.957513389897484801	0.007557448126411	0.00271177895685	0.00155428031626	0.0036647099791



```
          temperature_fahrenheit                   wind_speed_mph             HT_lag10_runs_scored            HT_lag10_runs_allowed 
                        1.129315                         6.680177                         4.201277                         1.477504 
                    HT_lag10_hits                  HT_lag10_errors             HT_lag10_left_on_base                HT_lag10_at_bats 
                        5.216203                         1.069140                         1.598684                         3.243891 
                HT_lag10_doubles                 HT_lag10_triples               HT_lag10_home_runs                    AT_lag10_wins 
                        1.591164                         1.121346                         2.056033                         3.687351 
            AT_lag10_runs_scored            AT_lag10_runs_allowed                    AT_lag10_hits                  AT_lag10_errors 
                        5.085248                         2.319071                         4.755153                         1.058162 
           AT_lag10_left_on_base                 AT_lag10_at_bats                 AT_lag10_doubles                 AT_lag10_triples 
                        1.579663                         3.147111                         1.626291                         1.102099 
              AT_lag10_home_runs          HT_lag01_win_percentage            HT_lag01_games_behind          AT_lag01_win_percentage 
                        2.071420                         2.573091                         2.249306                         2.496084 
           AT_lag01_games_behind                     lag01_totals                       coorsfield                        nightgame 
                        2.097634                         1.133048                         1.089845                         1.528243 
                           sunny                 wind_to_outfield               wind_from_outfield                wind_across_field 
                        2.618967                         4.986140                         4.916284                         5.867979 
              doubleheader_game1               doubleheader_game2  wind_speed_mph:wind_to_outfield wind_speed_mph:wind_from_outfield 
                        1.037380                         1.015326                        10.221854                         7.317931 
wind_speed_mph:wind_across_field                  nightgame:sunny 
                        9.745351                         2.924204 
```

11. DW Test Statistic

```
        Durbin-Watson test

data:  glm.bbdata.s1
DW = 2.0012, p-value = 0.4176
alternative hypothesis: true autocorrelation is greater than 0
```



```
          temperature_fahrenheit          wind_speed_mph            HT_lag10_runs_scored             HT_lag10_runs_allowed
                        1.129315                6.680177                        4.201277                          1.477504
                   HT_lag10_hits         HT_lag10_errors           HT_lag10_left_on_base                  HT_lag10_at_bats
                        5.216203                1.069140                        1.598684                          3.243891
                HT_lag10_doubles        HT_lag10_triples              HT_lag10_home_runs                     AT_lag10_wins
                        1.591164                1.121346                        2.056033                          3.687351
            AT_lag10_runs_scored   AT_lag10_runs_allowed                   AT_lag10_hits                   AT_lag10_errors
                        5.085248                2.319071                        4.755153                          1.058162
           AT_lag10_left_on_base        AT_lag10_at_bats                AT_lag10_doubles                  AT_lag10_triples
                        1.579663                3.147111                        1.626291                          1.102099
              AT_lag10_home_runs HT_lag01_win_percentage           HT_lag01_games_behind           AT_lag01_win_percentage
                        2.071420                2.573091                        2.249306                          2.496084
           AT_lag01_games_behind            lag01_totals                      coorsfield                         nightgame
                        2.097634                1.133048                        1.089845                          1.528243
                           sunny        wind_to_outfield              wind_from_outfield                 wind_across_field
                        2.618967                4.986140                        4.916284                          5.867979
              doubleheader_game1      doubleheader_game2 wind_speed_mph:wind_to_outfield wind_speed_mph:wind_from_outfield
                        1.037380                1.015326                       10.221854                          7.317931
wind_speed_mph:wind_across_field         nightgame:sunny
                        9.745351                2.924204
```

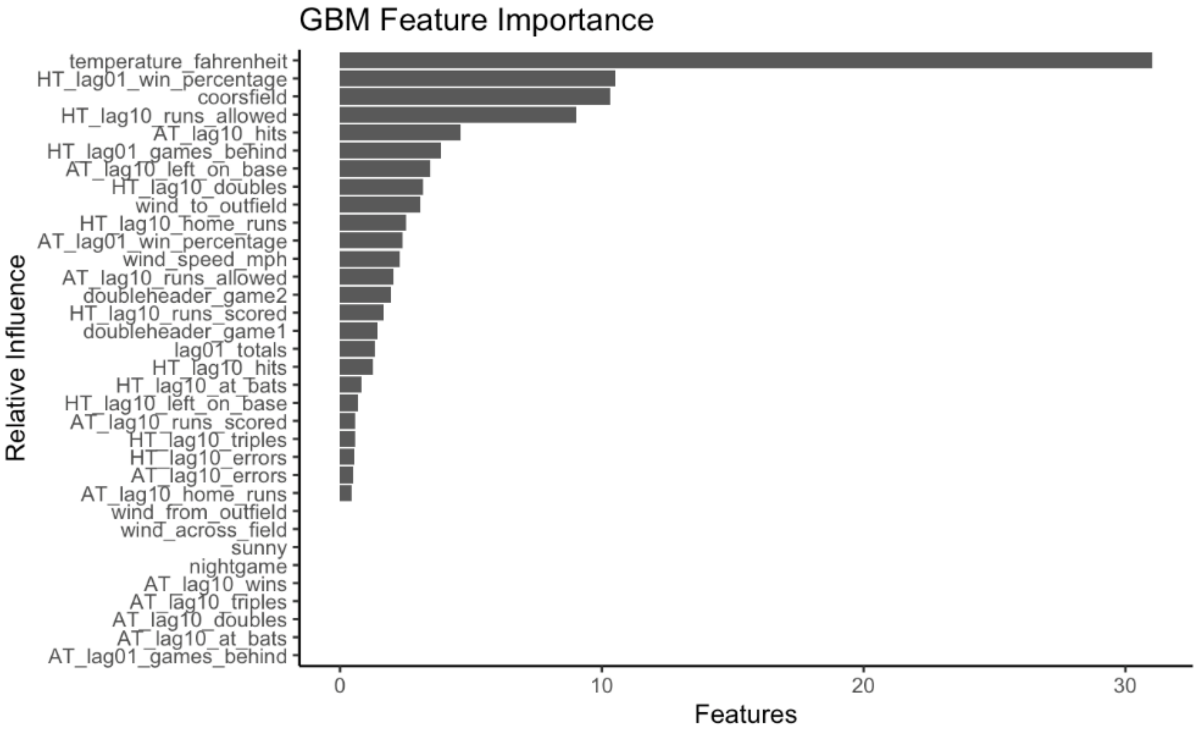
12. Models & Specifications

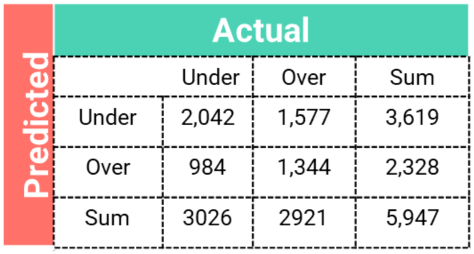
Model	Variable selection	Results
Logistic Regression	Business knowledge (LASSO for Feature selection)	53% accuracy at 0.49 cutoff
	PCA	45% Accuracy at 0.5 cutoff
Random Forest	Business knowledge	53% Accuracy
	PCA	53% Accuracy
Boosted Trees	Business knowledge	57% Accuracy
	PCA	55% Accuracy

LASSO Logistic Regression Confusion Matrix

```
Margins computed over dimensions
in the following order:
1: Predicted
2: Actual
         Actual
Predicted Under Over  sum
    Under  1617 1360 2977
    Over   1409 1561 2970
    sum    3026 2921 5947
```

13. Boosted Tree model (specification 1) Feature Importance & Confusion Matrix



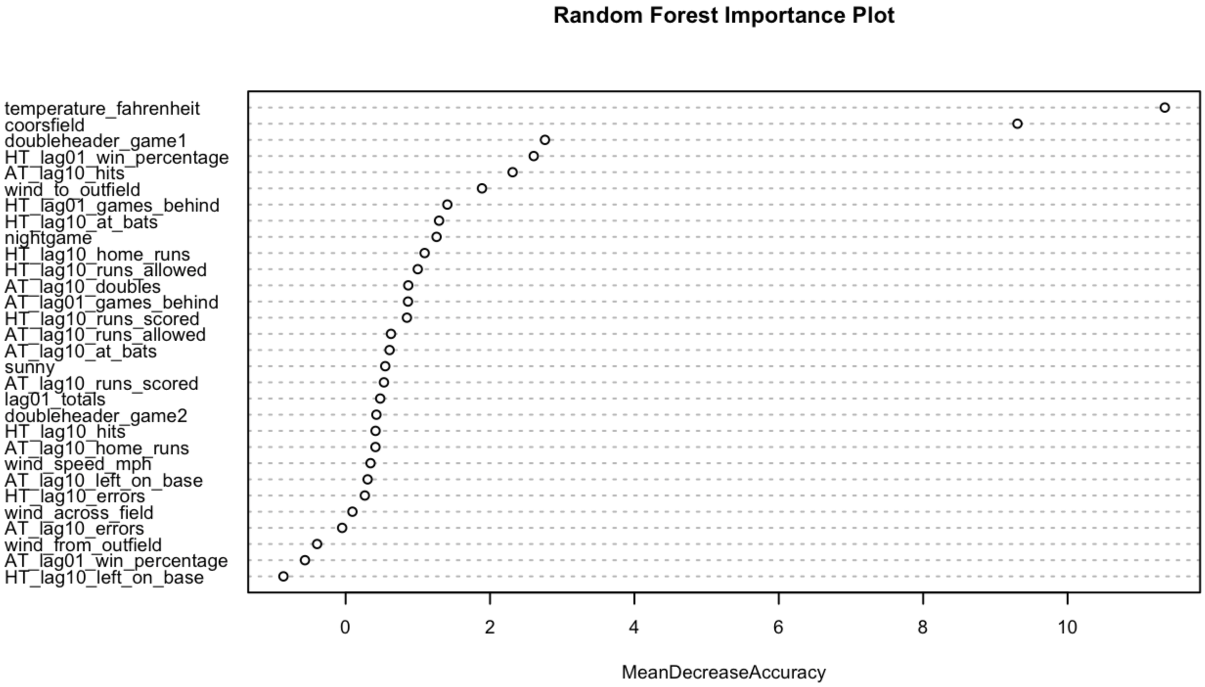


		Actual		
Predicted		Under	Over	Sum
	Under	2,042	1,577	3,619
	Over	984	1,344	2,328
	Sum	3026	2921	5,947

14. Accuracy, Error, Sensitivity, Specificity and False Positives for Boosted Model using All Predictors

```
     Accuracy Rate Error Rate Sensitivity Specificity False Positives
[1,]     0.5693627  0.4306373   0.4601164   0.6748182       0.3251818
```

15. Random Forest Features for GBM comparison



References

[1] https://www.supremecourt.gov/opinions/17pdf/16-476_dbfi.pdf

[2] https://www.sportico.com/business/sports-betting/2023/sports-betting-five-years-after- paspa-1234719556/

[3] https://frontofficesports.com/u-s-sports-betting-sets-records-in-2023-for-handle-revenue/

[4] https://www.washingtonpost.com/weather/2023/07/09/weather-baseball-homeruns- analytics-fantasy/#

[5] https://www.denver7.com/news/local-news/let-em-fly-coors-field-notoriously-known-as-a-home-run-park-heres-why

[6] https://sabr.org/sabermetrics/data

[7] https://docs.baseball.computer/#!/model/model.baseball_computer.game_start_info

[8] https://docs.baseball.computer/#!/model/model.baseball_computer.team_game_results

[9] https://docs.baseball.computer/#!/model/model.baseball_computer.standings

[10]https://docs.baseball.computer/#!/model/model.baseball_computer.team_game_offense_stats

Dataset