ROOT, ROOT, ROOT FOR THE HOME TEAM?

ESTIMATING THE PROBABILITY OF THE HOME TEAM WINNING A MLB GAME

Springboard DSCT

By Sinead O'Connor

December 2019

THE DATA

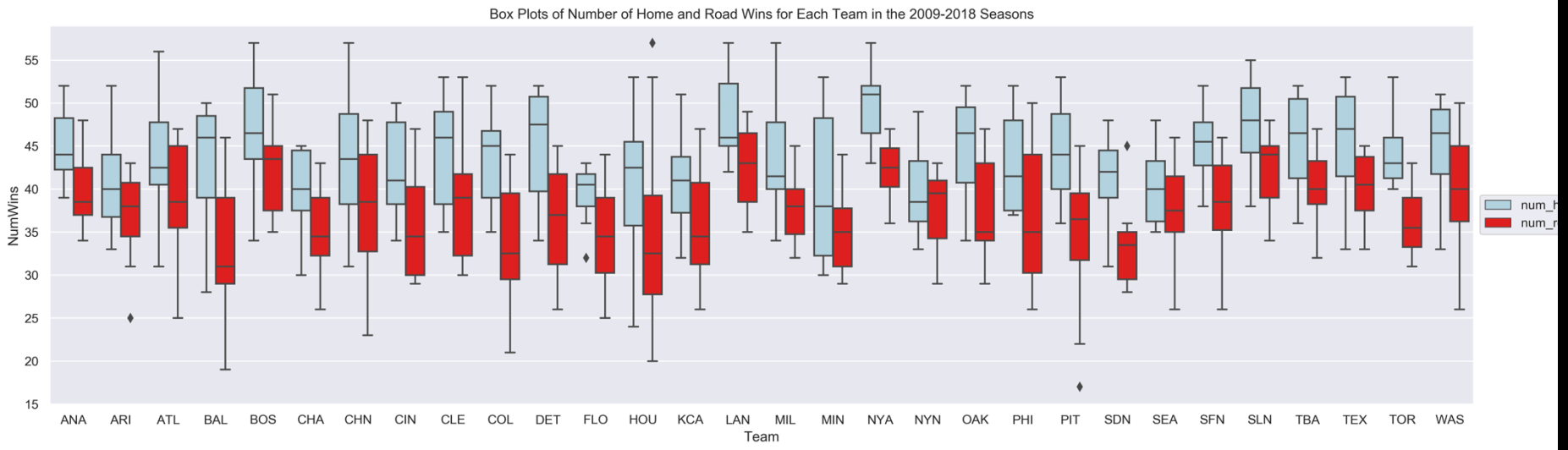
- **Game Logs for 2009-2018 seasons from Retrosheet.org**
  - Offensive, defensive, and pitching statistics on each game for both the home team and visiting team
  - General info, e.g. date, day of week, time of day (day/night), attendance
- **Starting Pitcher Logs for 2009-2018 seasons from BaseballMusings.com**
  - Starting pitcher stats for each individual game for each team, e.g. number of innings pitched, strikeouts, walks, and earned runs

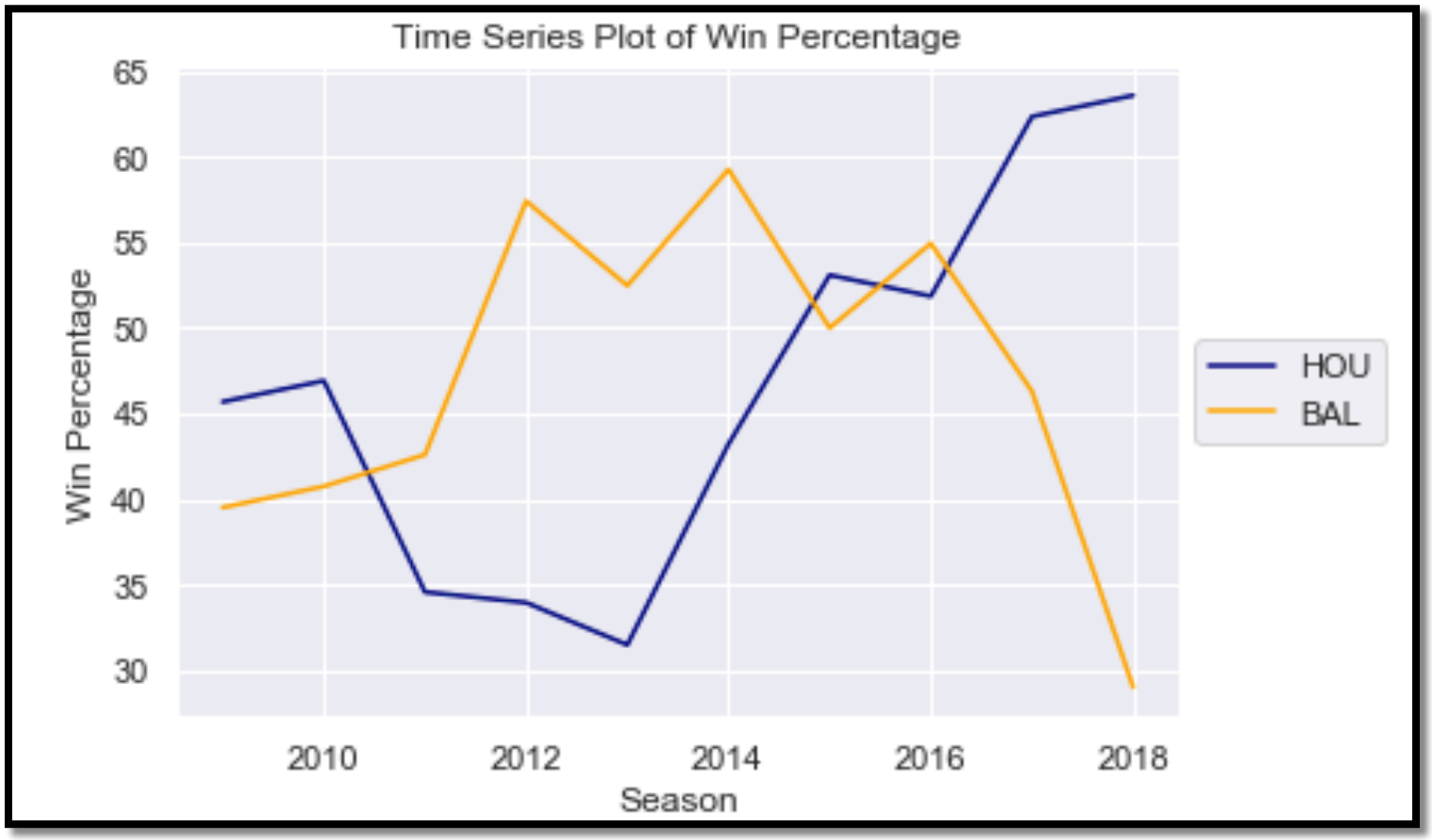
EXPLORATORY DATA ANALYSIS (EDA)

HOME/AWAY WINS

In the 2009-2018 seasons, the home team won 53.7% of games.

All 30 teams won more games at home than they did on the road.



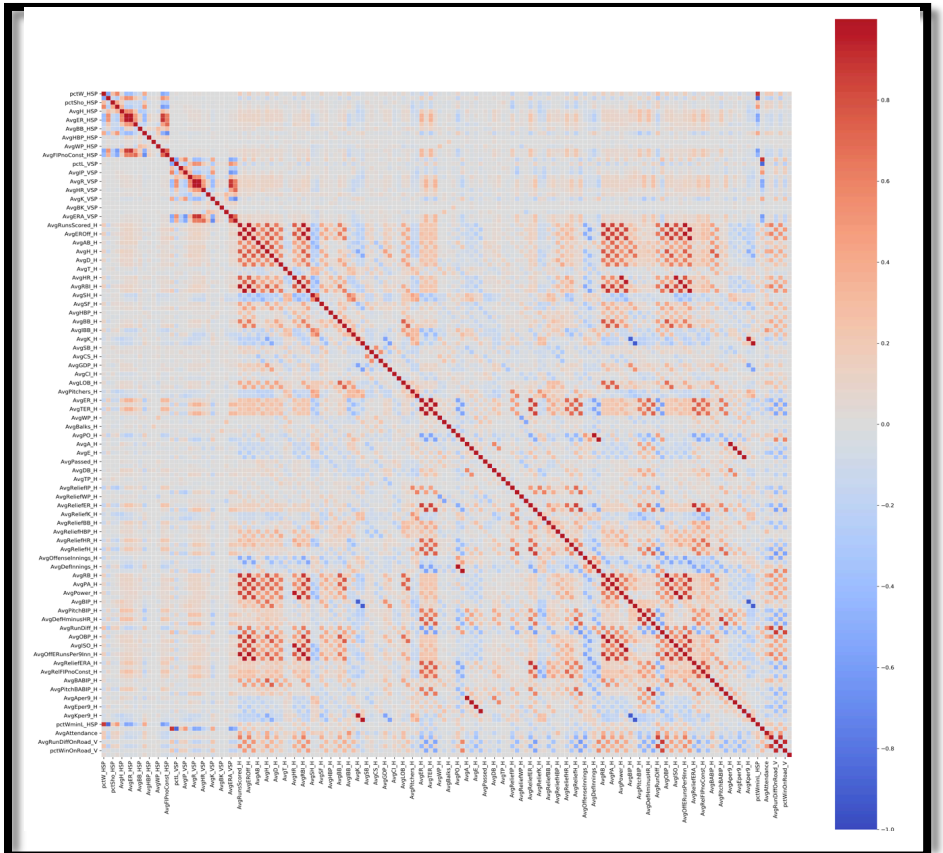


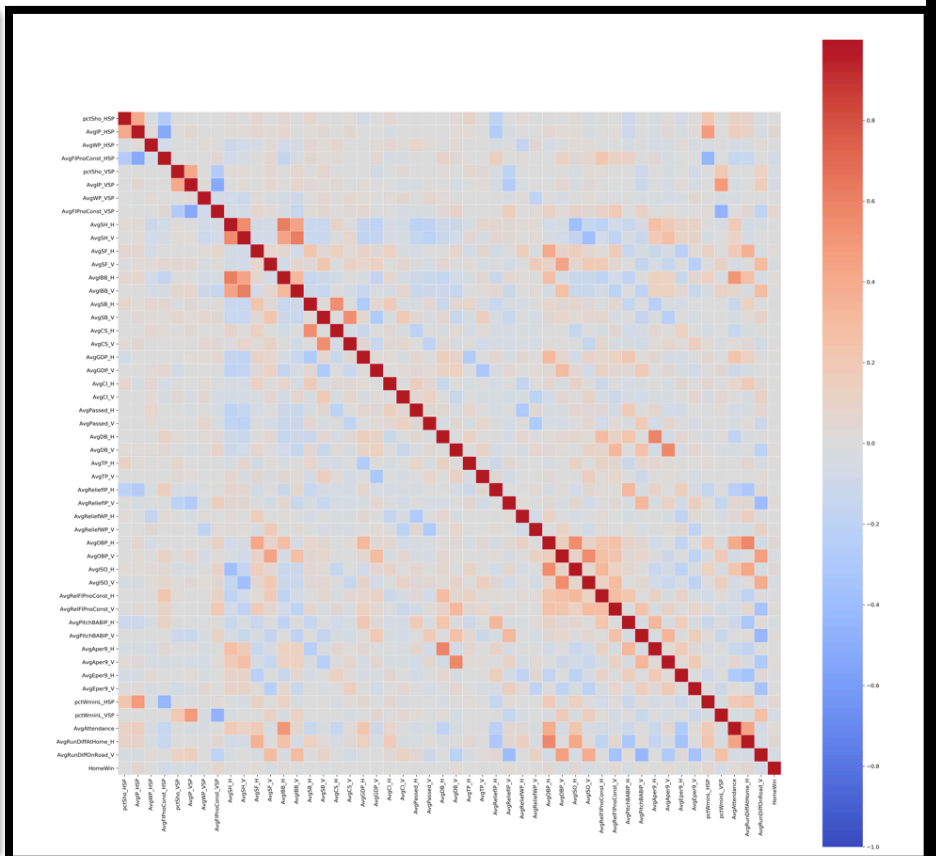
A team's performance one season does not necessarily mean they will have similar performance the following season,

MACHINE LEARNING

Baseline Logistic Regression Model

- For each, pair of collinear features, I removed one.
- Each vector represents one game with 49 features describing the home and away teams. Each feature was an expanding average beginning at the start of 2009 season.
- Target: 'HomeWin' column = 1 if the home team won, 0 if the home team lost the game





Baseline Logistic Regression Model: Results

[Training Classification Report:]

	precision	recall	f1-score	support
0	0.55	0.27	0.36	9948
1	0.56	0.81	0.66	11596
micro avg	0.56	0.56	0.56	21544
macro avg	0.55	0.54	0.51	21544
weighted avg	0.55	0.56	0.52	21544

Training Accuracy: 0.5585313776457482

[Test Classification Report:]

	precision	recall	f1-score	support
0	0.55	0.24	0.34	1148
1	0.55	0.82	0.66	1283
micro avg	0.55	0.55	0.55	2431
macro avg	0.55	0.53	0.50	2431
weighted avg	0.55	0.55	0.51	2431

Test Accuracy: 0.5475113122171946

AUC: 0.560206370630681

CONFUSION MATRIX:

		PREDICTED RESULT: LOSS	PREDICTED RESULT: WIN
TRUE RESULT	LOSS	279	869
	WIN	231	1052

EXTENDED MODELING:

RANDOM FOREST AND LOGISTIC REGRESSION

- **Created new features using rolling averages**
  - For stats pertaining to the whole team, I used the average of the previous 162 games.
  - For stats pertaining to the starting pitcher, I used the average of the pitcher's previous 30 starts.
  - For stats that only applied to the home team or visiting team (AvgAttendance, AvgRunDiffAtHome_H, and AvgRunDiffOnRoad_H), I calculated the averages of the previous 81 games.
- **Explicitly compared stats for home team and away team by subtracting visiting team stat from the corresponding stat of the home team**

RANDOM FOREST:

Hyper-parameter Tuning

Goal: To minimize out-of-bag (OOB) error rate

Parameters:
Max_features = log2
n_estimators = 1,000

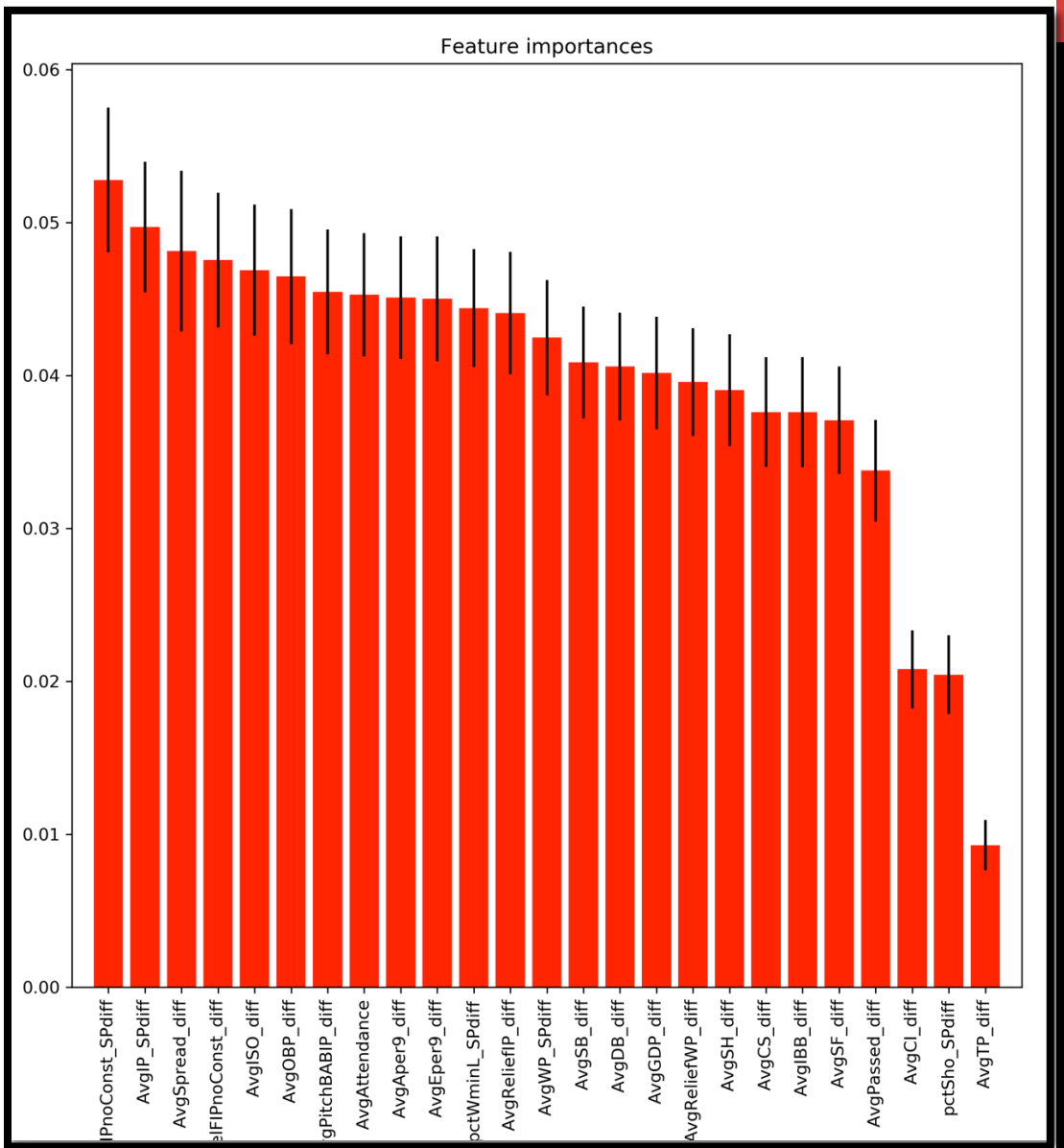
RANDOM FOREST: INITIAL RESULTS AND FEATURE IMPORTANCES

[Test Classification Report:]

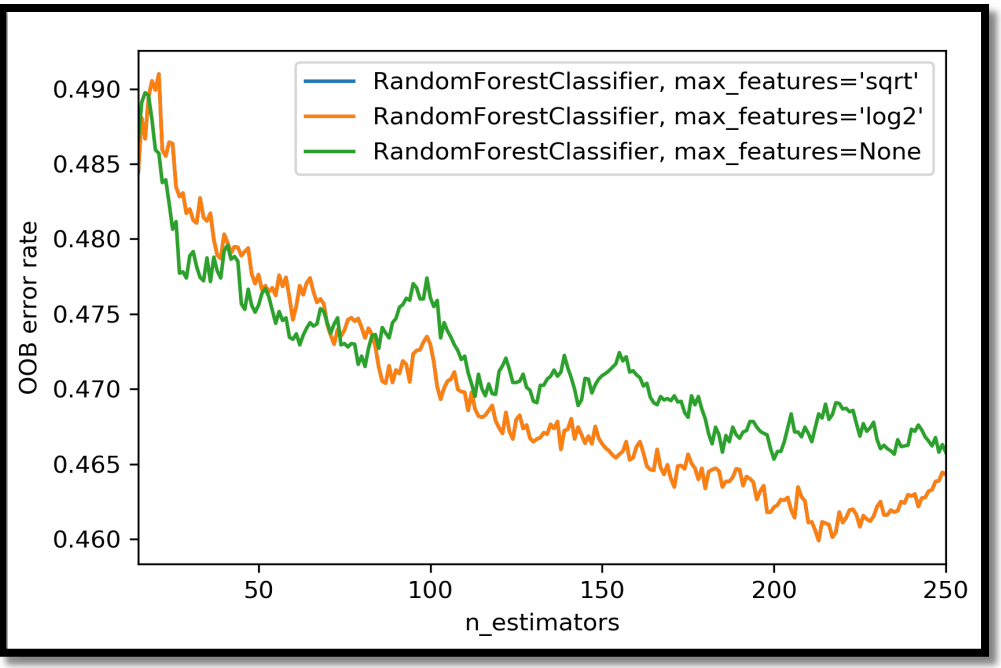
	precision	recall	f1-score	support
0	0.58	0.40	0.47	1148
1	0.58	0.75	0.65	1283
micro avg	0.58	0.58	0.58	2431
macro avg	0.58	0.57	0.56	2431
weighted avg	0.58	0.58	0.57	2431

Test Accuracy: 0.5808309337721103

AUC: 0.5989623079617947



RANDOM FOREST: REMOVING 6 LEAST IMPORTANT FEATURES



Parameters:

max_features = 'log2'

n_estimators = 1000

[Test Classification Report:]

	precision	recall	f1-score	support
0	0.59	0.42	0.49	1148
1	0.59	0.74	0.66	1283
micro avg	0.59	0.59	0.59	2431
macro avg	0.59	0.58	0.57	2431
weighted avg	0.59	0.59	0.58	2431

Test Accuracy: 0.5882352941176471

AUC: 0.6004736286089061

NOTE: Better results when we remove the 6 least important features

LOGISTIC REGRESSION RESULTS: USING SAME 19 FEATURES AS RANDOM FOREST

[Training Classification Report:]

	precision	recall	f1-score	support
0	0.52	0.56	0.54	9948
1	0.60	0.56	0.58	11596
micro avg	0.56	0.56	0.56	21544
macro avg	0.56	0.56	0.56	21544
weighted avg	0.56	0.56	0.56	21544

Training Accuracy: 0.5588098774600817

[Test Classification Report:]

	precision	recall	f1-score	support
0	0.56	0.59	0.57	1148
1	0.61	0.58	0.60	1283
micro avg	0.58	0.58	0.58	2431
macro avg	0.58	0.59	0.58	2431
weighted avg	0.59	0.58	0.59	2431

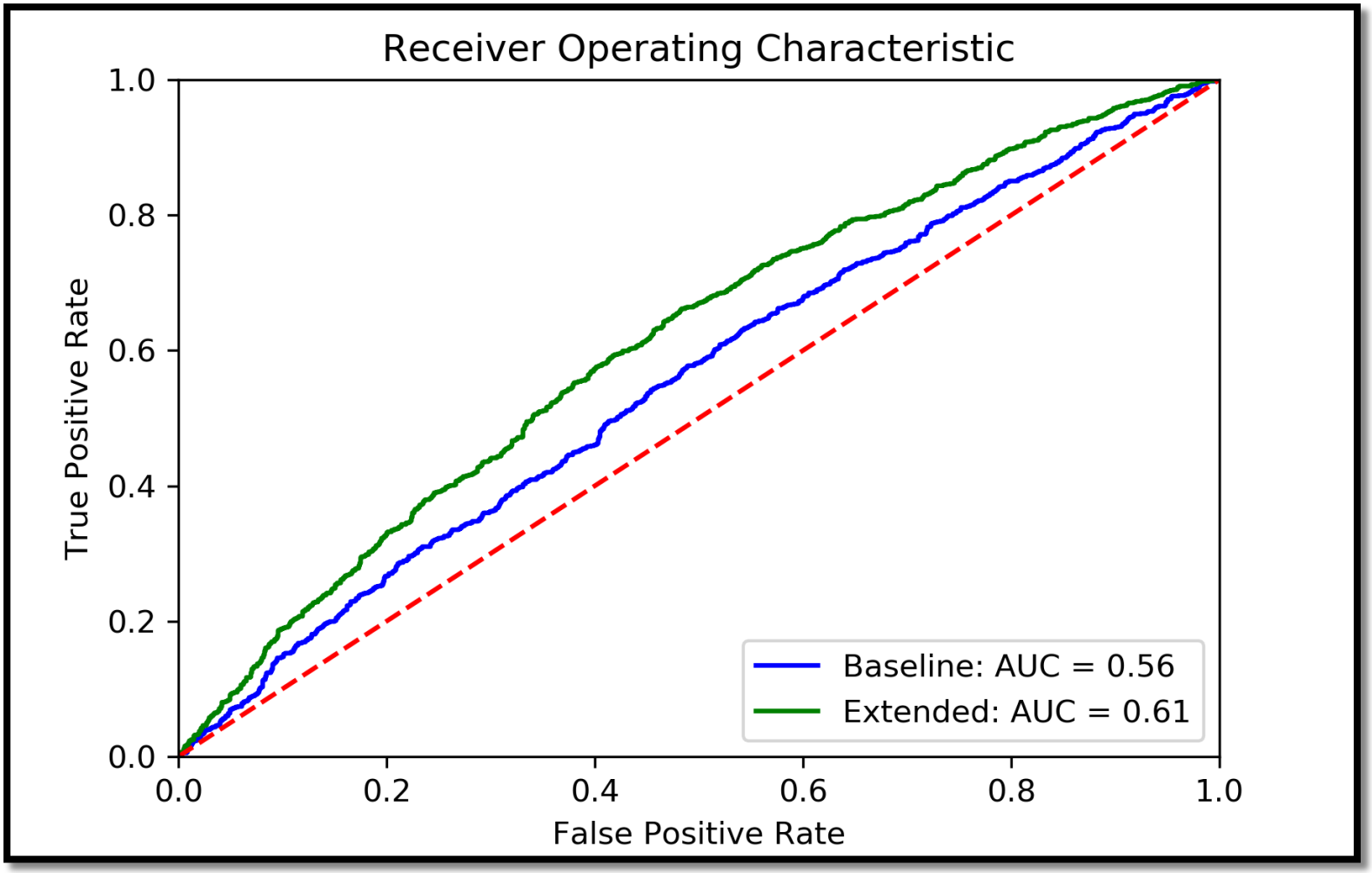
Test Accuracy: 0.5849444672974085

AUC: 0.6119891315269906

SELECTED FEATURES

1. AvgFIPnoConst_SPdiff
2. AvgIP_SPdiff
3. AvgRelFIPnoConst_diff
4. AvgSpread_diff = AvgRunDiffAtHome_H - AvgRunDiffOnRoad_V
5. AvgOBP_diff
6. AvgISO_diff
7. AvgAttendance (of home ball park)
8. AvgPitchBABIP_diff
9. AvgEper9_diff
10. AvgAper9_diff
11. AvgReliefIP_diff
12. pctWminL_SPdiff
13. AvgWP_SPdiff
14. AvgSB_diff
15. AvgDB_diff (DB = douple plays)
16. AvgGDP_diff (GDP = grounded into double play)
17. AvgReliefWP_diff
18. AvgSH_diff
19. AvgCS_diff

ROC CURVE


Receiver Operating Characteristic
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Baseline: AUC = 0.56
Extended: AUC = 0.61

SIMPLE SIMULATION

- Bet $100 dollars on a random set of 30 games in 2018 season
  - If correct – win $50
  - If incorrect – lose the $100

Date	Visiting Team	Home Team	Prob of Home Win LR	LR winnings	Prob of Home Win RF	RF winnings	Winner
2018-04-12	CHA	MIN	0.569380	50	0.629	50	Home
2018-04-13	MIL	NYN	0.457341	-100	0.554	50	Home
2018-04-13	SFN	SDN	0.501470	50	0.553	50	Home
2018-04-19	TOR	NYA	0.568426	50	0.699	50	Home
2018-04-30	NYA	HOU	0.486090	-100	0.561	50	Home
2018-05-01	NYA	HOU	0.515514	-100	0.561	-100	Visitor
2018-05-04	CHN	SLN	0.481290	-100	0.414	-100	Home
2018-05-15	CLE	DET	0.399107	-100	0.494	-100	Home
2018-05-20	LAN	WAS	0.491341	50	0.492	50	Visitor
2018-05-21	FLO	NYN	0.504208	50	0.512	50	Home
2018-05-21	KCA	SLN	0.578380	50	0.633	50	Home
2018-05-29	CHA	CLE	0.660552	50	0.682	50	Home
2018-06-02	TOR	DET	0.438834	-100	0.503	50	Home
2018-06-18	FLO	SFN	0.536889	-100	0.573	-100	Visitor
2018-06-19	TEX	KCA	0.466499	50	0.554	-100	Visitor
2018-06-29	KCA	SEA	0.584006	50	0.528	50	Home
2018-07-05	ATL	MIL	0.535343	50	0.535	50	Home

LogReg Model:

Winnings: -600

ROI: -20%

RF Model:

Winnings: $150

ROI: 5%

SIMPLE SIMULATION (CONT.)

If betting $100 on all 2,431 games of the 2018 season:

Test Metric	Using LR Predictions	Using RF Predictions
Winnings	-$29,800	-$28,600
ROI	-12.25%	-11.76%

Ensemble Method: If only betting on the 1917 games in the 2018 season for which BOTH models predicted the home team would win or BOTH predicted the home team would lose:

Test Metric	Value
Winnings	-$16,350
ROI	-8.53%
Accuracy	60.98%

COMPARING RESULTS

CONFUSION MATRIX FOR LR:

		PREDICTED RESULT: LOSS	PREDICTED RESULT: WIN
TRUE RESULT	LOSS	677	471
	WIN	538	745

CONFUSION MATRIX FOR RF:

		PREDICTED RESULT: LOSS	PREDICTED RESULT: WIN
TRUE RESULT	LOSS	479	669
	WIN	332	951

Test Metric	Chosen Model	Value of Metric	Business Interpretation
Accuracy	RF	58.82%	The more accurate, the better, and the more money you will make using the model to bet. However, accuracy could be misleading, for example if the model predicted that the home team won every game it would still have an accuracy of about 53%,
Precision	Either	0.59	The RF had 0.59 precision for both classes while the LR had 0.56 precision for class 0 and 0.61 precision for class 1. This means when the model predicted the visiting team won it was correct 56% of the time and when it predicted the home team won it was right 61% of the time.
Recall	RF	0.59	Choosing based on the model with higher avg recall may be misleading. The Random Forest's recall for class 0 (home loss) is 0.42 , while its recall for class 1 is 0.74, which means when the home team actually wins the model is correct 74% of the time, but when the visiting team actually wins it's only correct 42% . The LR model has recalls of 0.59 and 0.58 for class 0 and 1.
F1 Score	LR	0.59	This is a good balance between precision and recall.
AUC	LR	0.6119	The higher the AUC, the better. A high AUC indicates that the ratio of true positive rate (recall) and false positive rate (predicting that the home team won when they actually lost) is higher.
Log Loss	LR	0.674	We want to minimize log loss. Log loss penalizes highly confident predictions that are wrong. Even if a client loses money they maybe more upset if a model predicts that the home team will win with 90% probability and they lose than if the model had predicted 52% for example, and the client knew it was a riskier bet.
Winnings/ ROI	RF	-$28,600/ -11.76%	People using the service want to maximize winning or at least minimize losses.

RECOMMENDATIONS TO THE CLIENT

- **If the person is using the client's service is solely winning money based whether they correctly predicting the winner of the game, I would recommend that bet based on the predictions of the Random Forest model, as it has the highest accuracy and the losses are minimized.**
- **I would also suggest that they bet on games in which both the Logistic Regression and the Random Forest model predict the same outcome, because in the 1,917 games in which this was the case, it was 60.9% accurate.**
- **If the person using the client's service plans to bet more or less money based on the predicted probabilities, I would suggest that they use the results of the Logistic Regression model, because it has a slightly lower log loss.**
- **More work to be done before 2020 season starts**…