JOSEPH DARBY

MATCH PREDICTIONS

EUROPEAN “FOOTBALL” CHALLENGE

Leave it all on the pitch

DATA SCIENCE PROCESS

DEFINE THE PROBLEM

GATHER THE DATA

DATA EXPLORATION

MODEL THE DATA

EVALUATE MODEL

ANSWER THE PROBLEM

PROBLEM STATEMENT

Using the Soccer Database provided, can we build a multi-classification model that predicts the outcome of a given match with better accuracy than the bookies, with majority class distribution percentage of 46%?

DATA SCIENCE PROCESS

DEFINE THE PROBLEM

GATHER THE DATA

DATA EXPLORATION

MODEL THE DATA

EVALUATE MODEL

ANSWER THE PROBLEM

ABOUT THE DATABASE

THE TOP LEAGUES FROM

11 COUNTRIES

DATA ON

25,000+ MATCHES

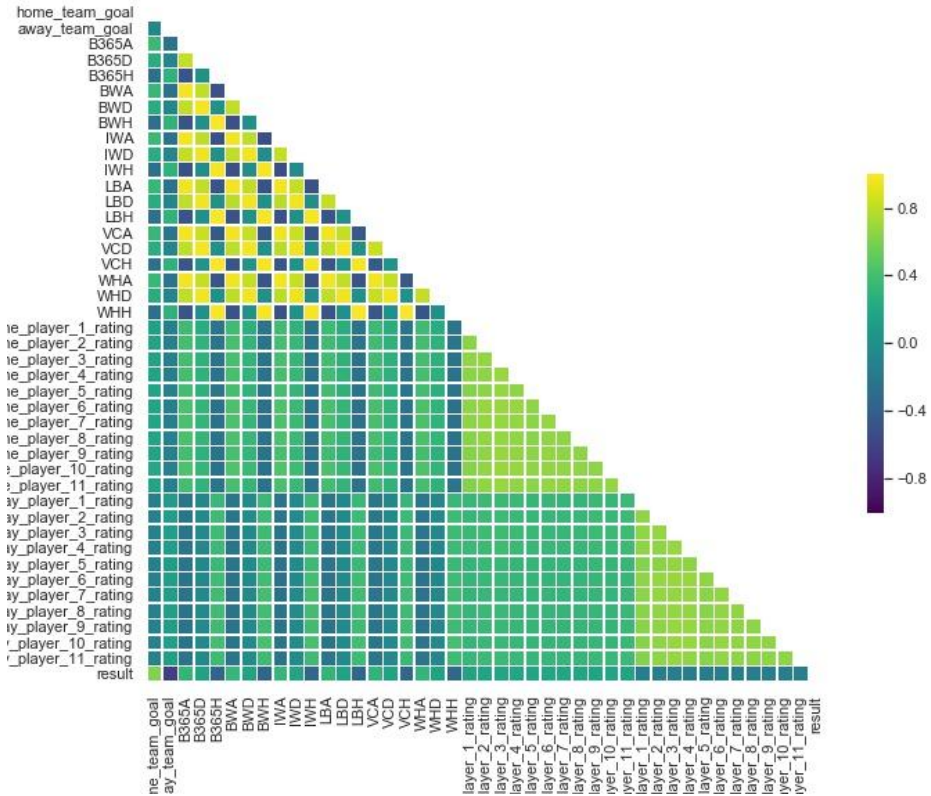
&

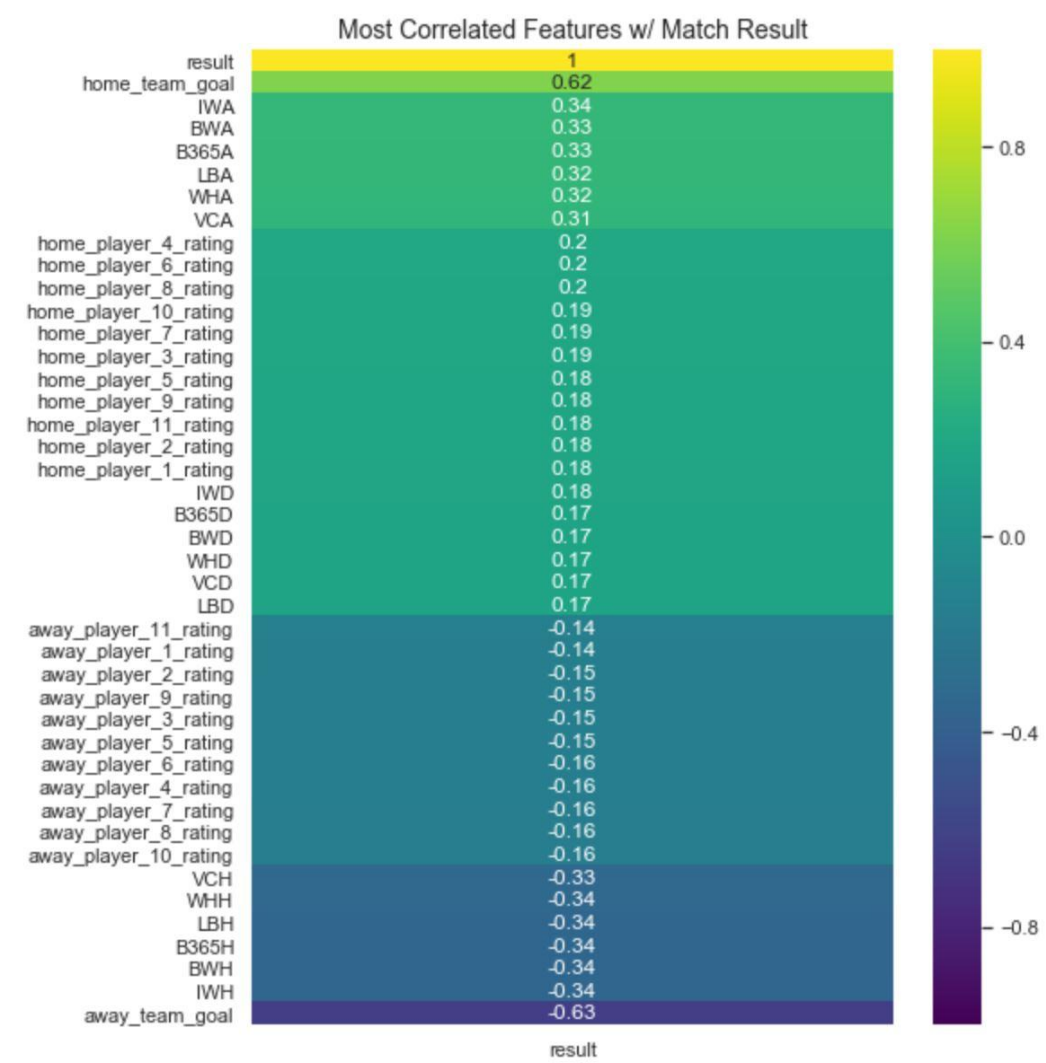
10,000+ PLAYERS & . . .

ABOUT THE DATABASE

- **PLAYER & TEAM ATTRIBUTES** SOURCED FROM EA SPORTS FIFA VIDEO GAME WITH WEEKLY UPDATES
- **TEAM LINEUP & FORMATION** IN X AND Y COORDINATES....?
- **BETTING ODDS** FROM 10 BETTING PROVIDERS
- **MATCH EVENT DETAILS** FOR 10,000+ MATCHES, SUCH AS GOALS, POSSESSION PERCENTAGE, CARDS, AND CORNER KICKS

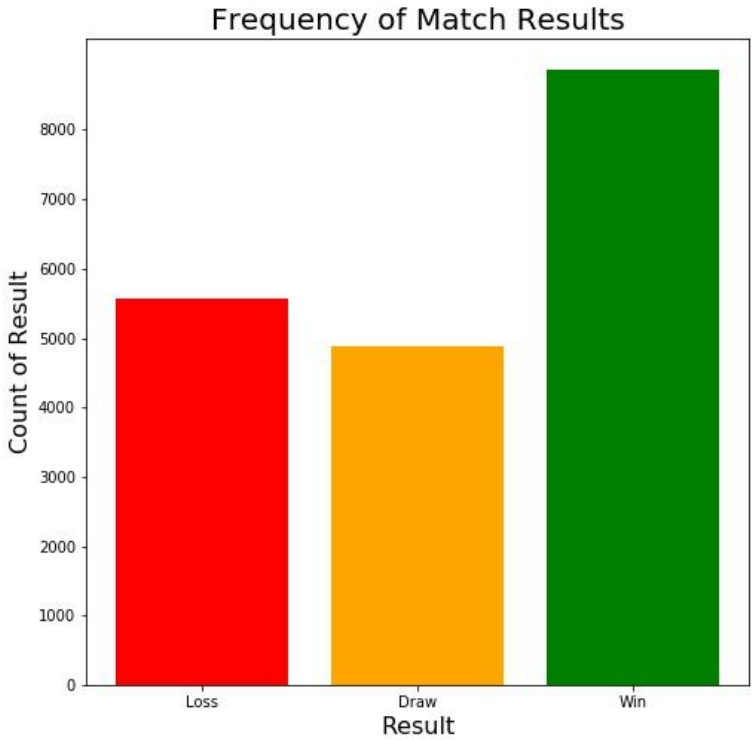
OVERALL CORRELATION



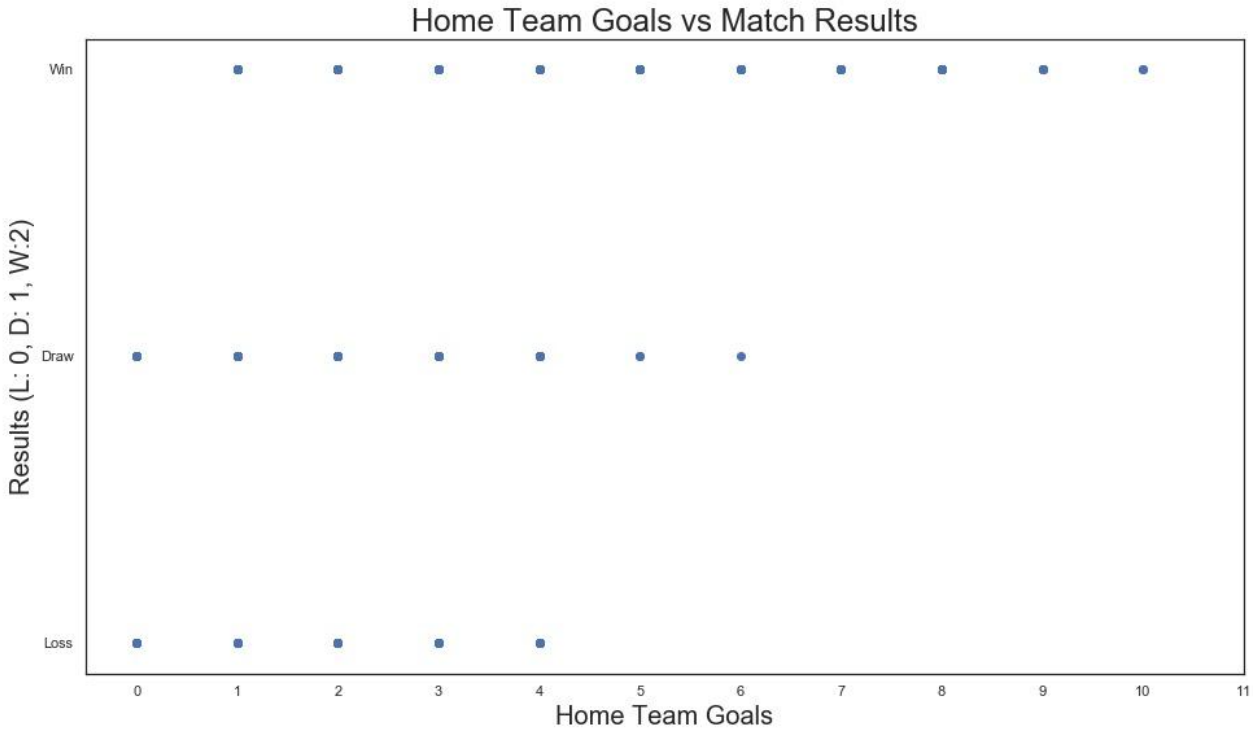


TARGET VECTOR CORRELATION

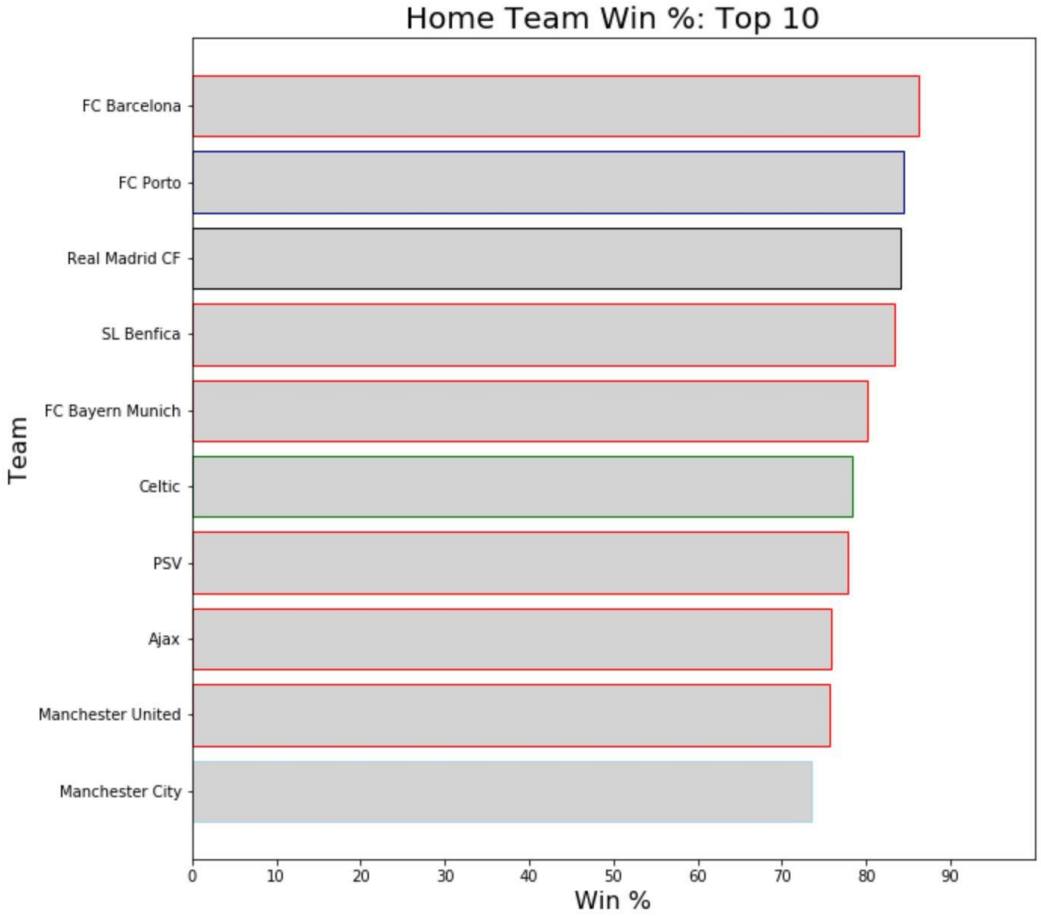
HOW DOES THE HOME TEAM PERFORM?



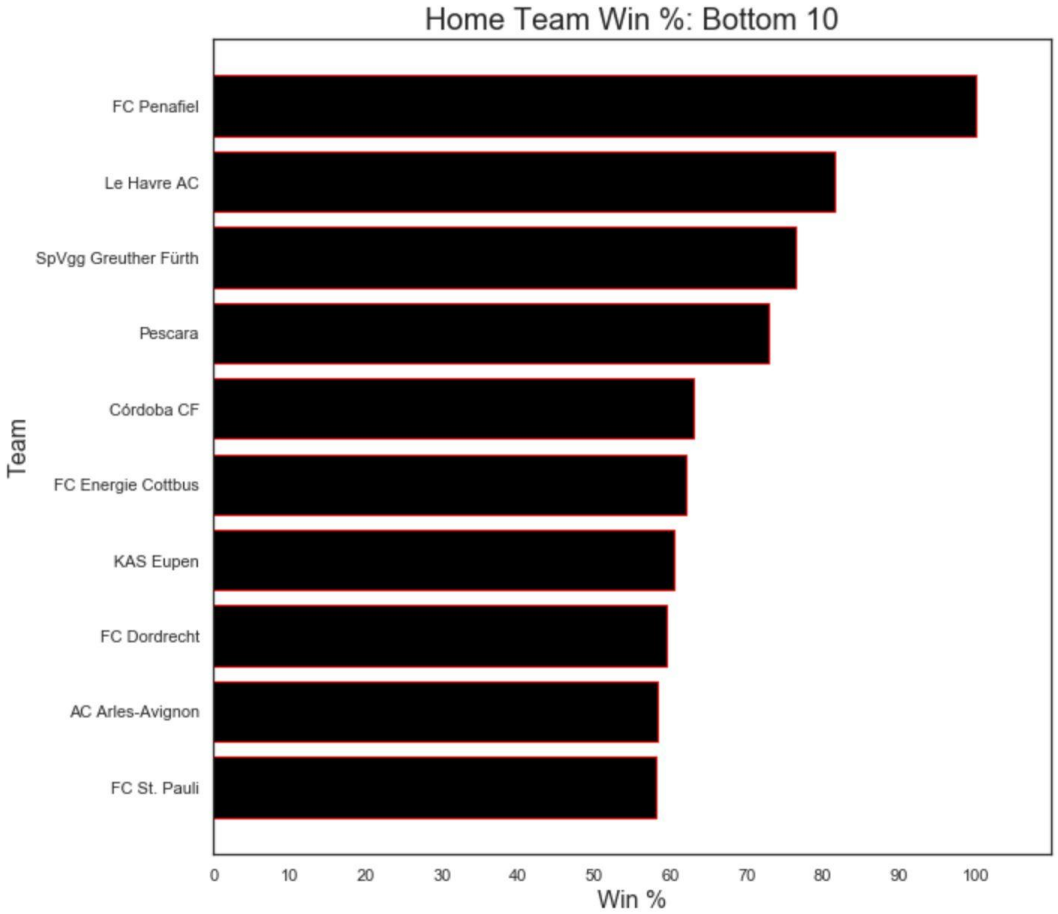
HOW DOES THE NUMBER OF GOALS CORRELATE TO MATCH RESULTS?



TOP 10 HOME TEAMS



BOTTOM 10 HOME TEAMS



✋ PROBLEMS, PROBLEMS, EVERYWHERE!!!

- 11 **DIFFERENT SOURCES** OF DATA BETWEEN LEAGUES
- 25,000 MATCHES…YES!

  MANY DON'T HAVE MATCH EVENT DETAILS…NOOOOO
- 10,000 PLAYERS,

  BUT **NO COMMON KEY** TO JOIN THEM WITH THEIR TEAM
- X AND Y FIELD COORDINATES FOR PLAYER POSITION?…JUST PUT THE POSITION NAME IN A STRING AND I'LL **DUMMY THEM!**
- 10 BET BROKERS, YET **SO MANY NULLS**
- ALL MATCH DETAILS **WRAPPED IN HTML!** LORD HELP ME!

WE'RE GOING FAST & CHEAP

- **7 TABLES** OF DATA**: used SQLite3 for querying**
  - TEAM & TEAM_ATTRIBUTES
  - PLAYERS & PLAYER_ATTRIBUTES
  - COUNTRY & LEAGUE
  - MATCH TABLE
- REMOVED **ALL** NULLS**:** SHAPE -> **(1762, 311)**
  - ORIGINAL DATAFRAME: **(25,979, 119)**
- DECIDED TO MODEL **TWO DIFFERENT SUBSETS** OF THE DATA**:**
  - REMOVED **ALL NULLS** (ABOVE)
  - **SELECTION OF FEATURES:** MOST RELEVANT
    - TWO REPUTABLE BET SITES: **BET365 & WILLIAM HILL**
- **MODELS: LOGISTIC REGRESSION CLASSIFIER** & **RANDOM FOREST CLASSIFIER**

DATA SCIENCE PROCESS

DEFINE THE PROBLEM

GATHER THE DATA

DATA EXPLORATION

MODEL THE DATA

EVALUATE MODEL

ANSWER THE PROBLEM

PURE DATAFRAME
LOGISTIC REGRESSION

SHAPE	1762 rows, 311 series

CLASSES

HOME LOSS (0)	0.284336
HOME DRAW (1)	0.259932
HOME WIN (2)	0.455732

LOGISTIC REGRESSION
QUICK WORKFLOW

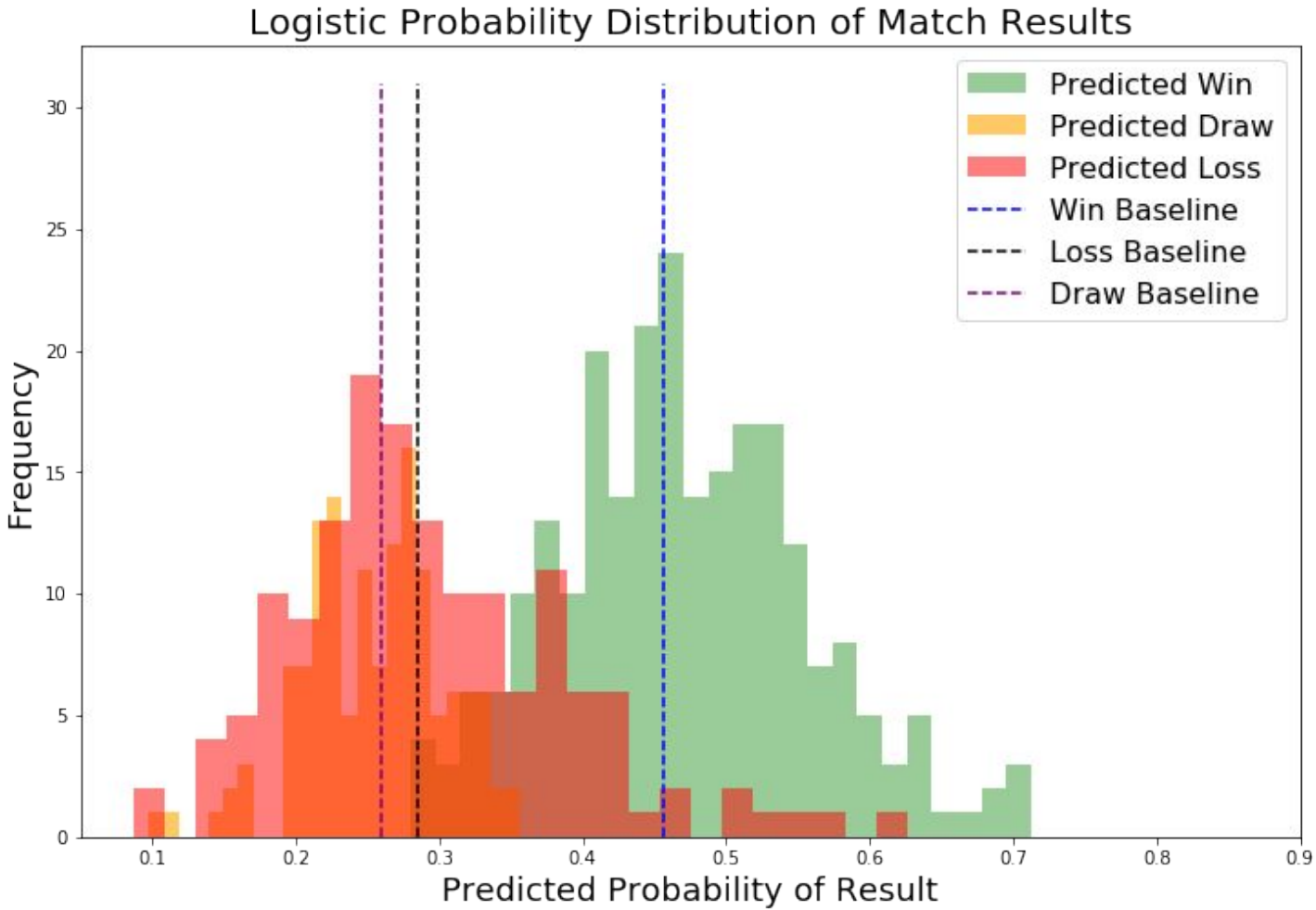
- **Set up X & Y:** dropped home and away goals
- **Train/test/split:** 30% test group; Stratify y
- **Instantiate model:** multi_class='multinomial
- **CV Scores, Fit, Predict(X_test), Evaluate Predictions**

CV TRAIN SCORE	0.4461
CV TEST SCORE	0.4329
ACCURACY	0.4442

Majority Class Distribution is better, BASELINE = 45.57%

LOGISTIC REGRESSION
EVALUATION

	true_values	pred_probs	model_preds	model_proba
1305	2	0.449876	2	0.449876
302	2	0.306967	1	0.399092
181	0	0.279558	2	0.453942
1700	1	0.224527	0	0.430904
124	0	0.236071	2	0.449106


Logistic Probability Distribution of Match Results
Predicted Win
Predicted Draw
Predicted Loss
Win Baseline
Loss Baseline
Draw Baseline
Frequency
0
5
10
15
20
25
30
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Predicted Probability of Result

PURE DATAFRAME
RANDOM FOREST

SHAPE	1762 rows, 311 series

CLASSES

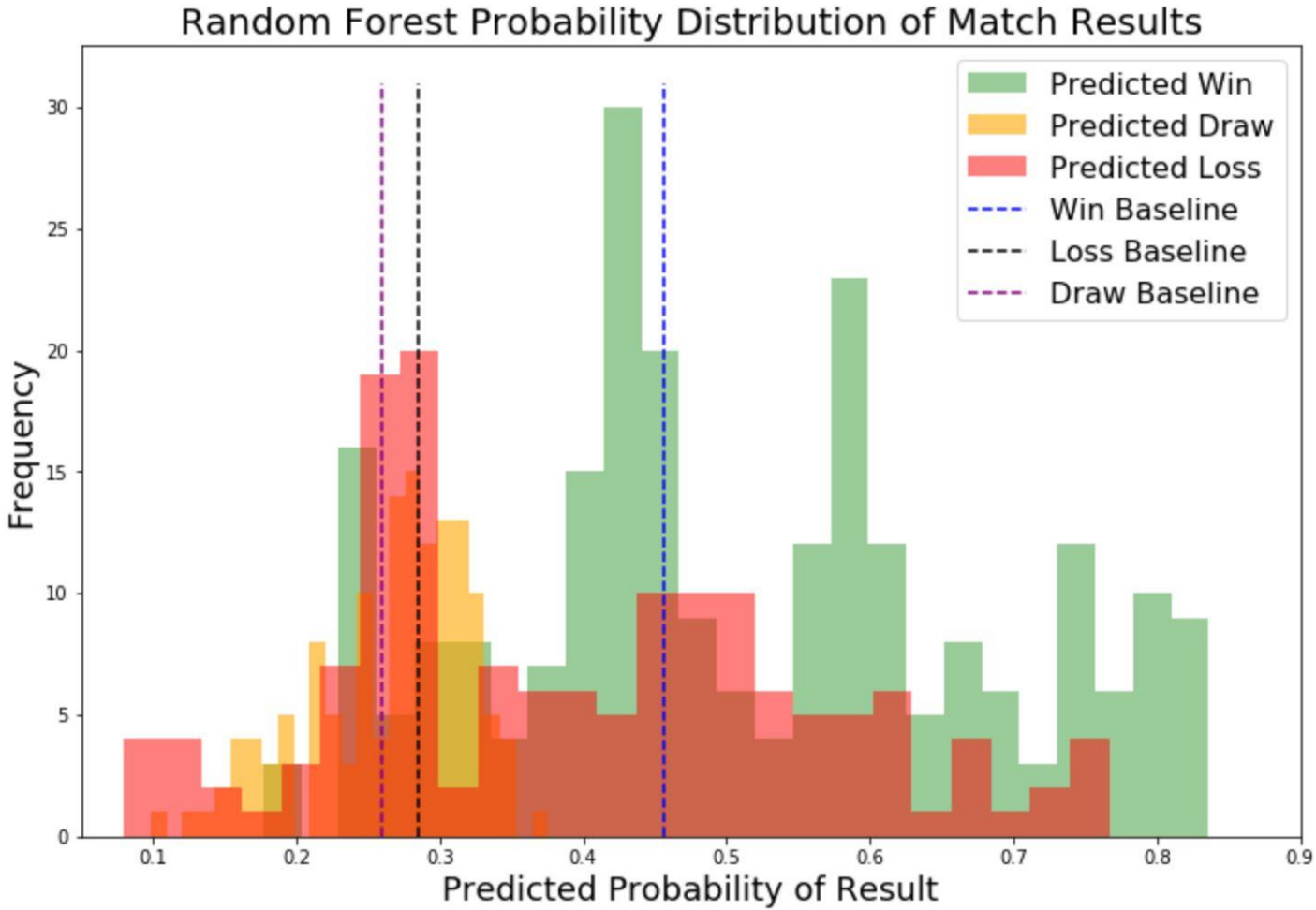
HOME LOSS (0)	0.284336
HOME DRAW (1)	0.259932
HOME WIN (2)	0.455732

RANDOM FOREST
QUICK WORKFLOW

- **Set up X & Y:** dropped home and away goals
- **Train/test/split:** 30% test group; Stratify y
- **Instantiate model:**
- **CV Scores, Fit, Predict(X_test), Evaluate Predictions**

5-CV TRAIN SCORE	0.4810
GS TRAIN SCORE	0.5539
GS TEST SCORE	0.5331

BEAT THE BASELINE = 45.57% (~7%)


Random Forest Probability Distribution of Match Results
Predicted Win
Predicted Draw
Predicted Loss
Win Baseline
Loss Baseline
Draw Baseline
Frequency
0
5
10
15
20
25
30
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Predicted Probability of Result

PARSED DATAFRAME
LOGISTIC REGRESSION

SHAPE	22,568 rows, 549 series

CLASSES

HOME LOSS (0)	0.2878
HOME DRAW (1)	0.2531
HOME WIN (2)	0.4591

LOGISTIC REGRESSION
QUICK WORKFLOW

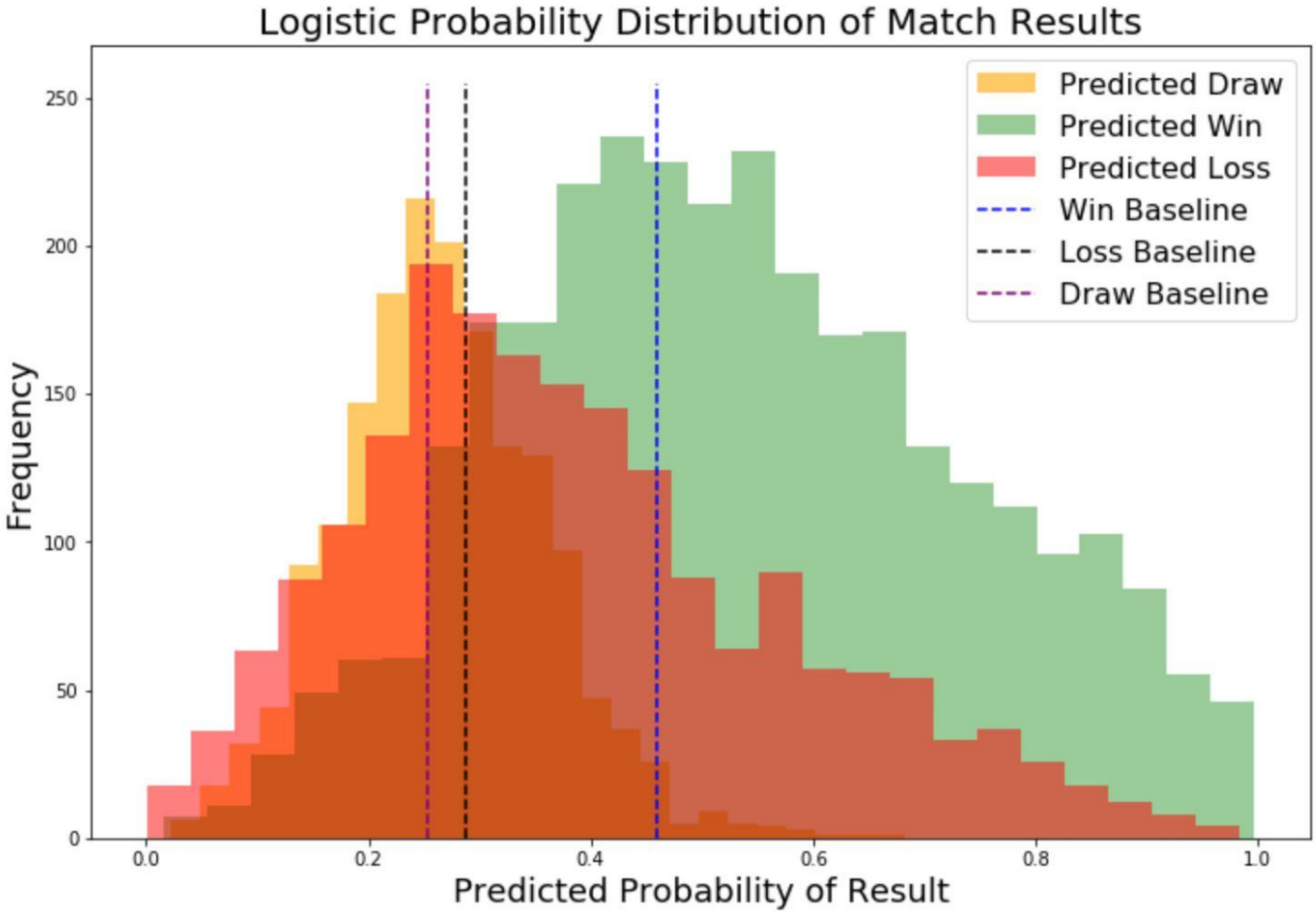
- **Set up X & Y:** dropped home and away goals
- **Train/test/split:** 30% test group; Stratify y
- **Instantiate model:** multi_class = 'multinomial', max_iter = 2000
- **CV Scores, Fit, Predict(X_test), Evaluate Predictions**

CV TRAIN SCORE	0.5047
CV TEST SCORE	0.4844
ACCURACY	0.5134

** BASELINE = 51.34%**

RANDOM FOREST
EVALUATION

	true_values	pred_probs	model_preds	model_proba
18715	1	0.343823	2	0.451362
10380	1	0.066851	2	0.897474
10108	1	0.359173	0	0.464561
7185	0	0.243779	2	0.513100
9544	0	0.373625	2	0.421978


Logistic Probability Distribution of Match Results
Predicted Draw
Predicted Win
Predicted Loss
Win Baseline
Loss Baseline
Draw Baseline
Frequency
250
200
150
100
50
0
0.0
0.2
0.4
0.6
0.8
1.0
Predicted Probability of Result

PARSED DATAFRAME
RANDOM FOREST

SHAPE	22,568 rows, 549 series

CLASSES

HOME LOSS (0)	0.2878
HOME DRAW (1)	0.2531
HOME WIN (2)	0.4591

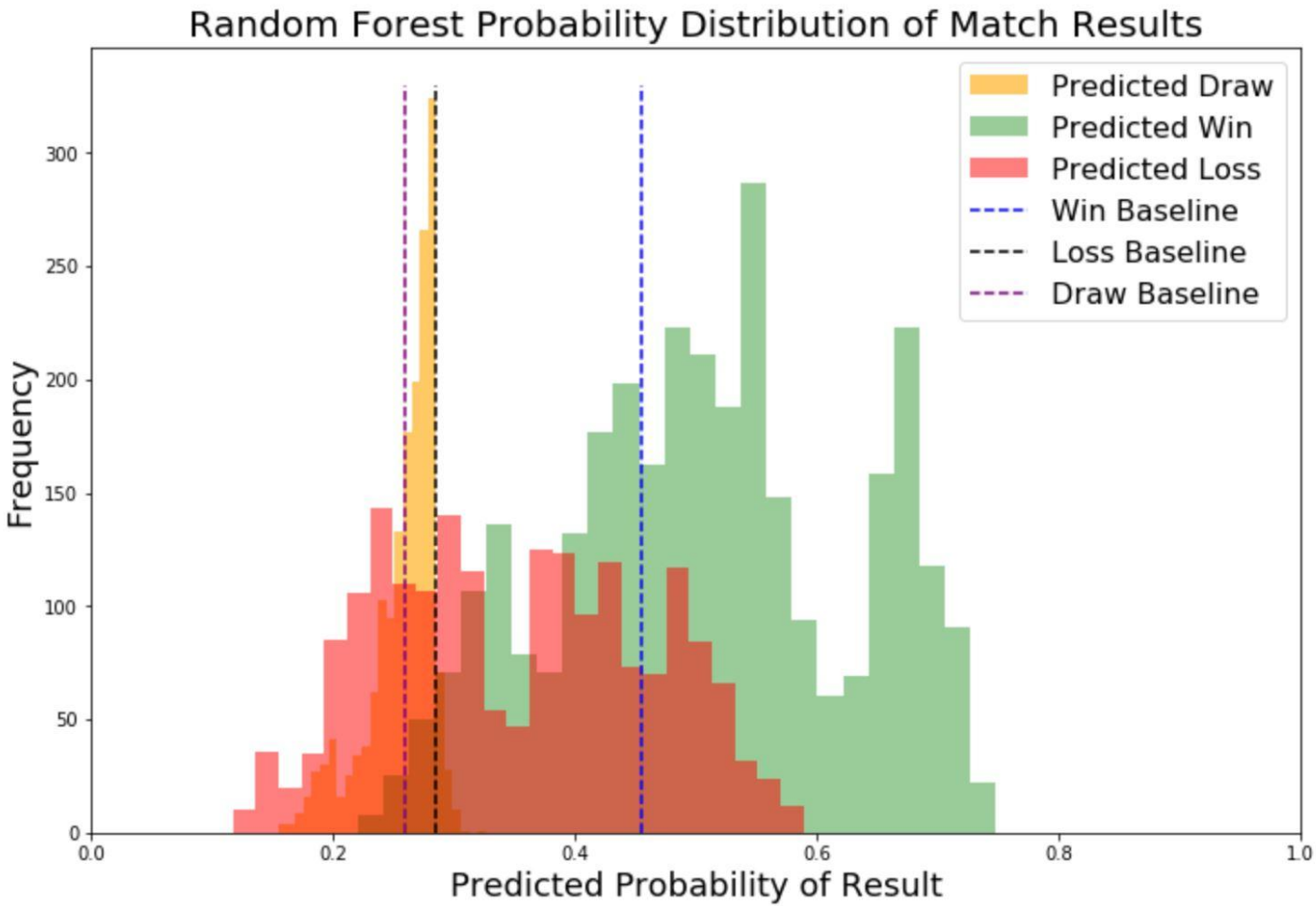
RANDOM FOREST
QUICK WORKFLOW

- **Set up X & Y:** dropped home and away goals
- **Train/test/split:** 30% test group; Stratify y
- **Instantiate model:**
- **CV Scores, Fit, Predict(X_test), Evaluate Predictions**

5-CV TRAIN SCORE	0.4950
GS TRAIN SCORE	0.5350
GS TEST SCORE	0.5322

RANDOM FOREST
EVALUATION

	true_values	pred_probs	model_preds	model_proba
18715	1	0.275922	2	0.493343
10380	1	0.179491	2	0.685453
10108	1	0.273887	0	0.397937
7185	0	0.269190	2	0.453119
9544	0	0.381091	0	0.381091


Random Forest Probability Distribution of Match Results
Predicted Draw
Predicted Win
Predicted Loss
Win Baseline
Loss Baseline
Draw Baseline
Frequency
0
50
100
150
200
250
300
0.0
0.2
0.4
0.6
0.8
1.0
Predicted Probability of Result

DATA SCIENCE PROCESS

DEFINE THE PROBLEM

GATHER THE DATA

DATA EXPLORATION

MODEL THE DATA

EVALUATE MODEL

ANSWER THE PROBLEM

G  AL!

VERDICT:
IT CAN BE DONE!!!
but, improved upon...