Euro Leagues Soccer Outcome Prediction

Luca Petrelli, Ph.D.

Outline

- Problem Description
- Data
- Data Exploration
- Features
- Classifiers
- Log Loss and Accuracy Evaluation
- Probability Calibration
- Return On Investment
- Conclusion
- Recommendations

Problem Description

Predicting soccer outcomes for games in the 5 major European leagues. Setting probabilities of Home win, Draw, or Away win. Predicting games outcomes is a difficult process, and bookmakers would like to improve their predictions.

Goal: improve on current bookmakers' odds assignment. Deliver a model that does better, not only in terms of accuracy but also in terms of actual returns for the bookmakers.

Data

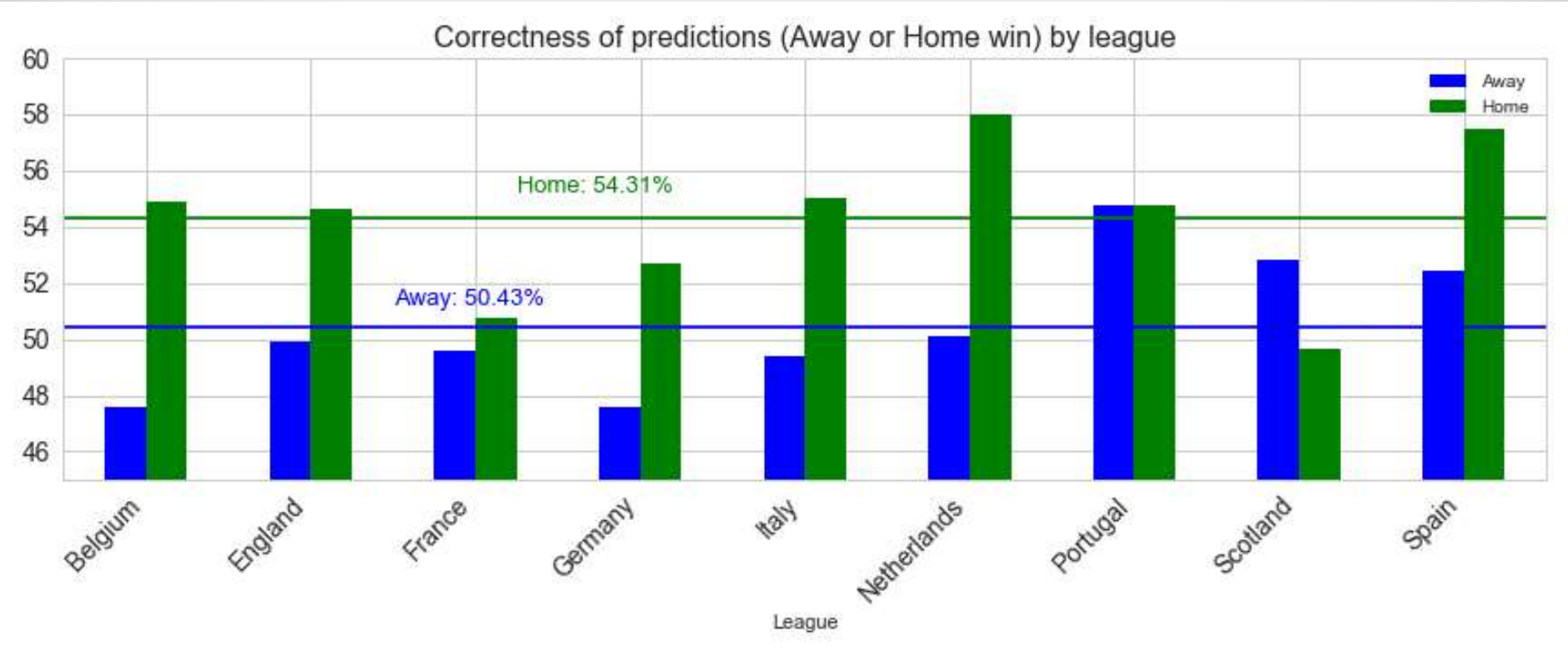
Data is taken from the European Soccer Database at:
https://www.kaggle.com/hugomathien/soccer
8 seasons of soccer statistics from 11 European Leagues, 4 tables:

- Player: names, birthdates
- Player_Stats: FIFA games attributes
- Team: short and long names, team_ids
- Match: game scores, dates, in-game events, bookmakers' odds

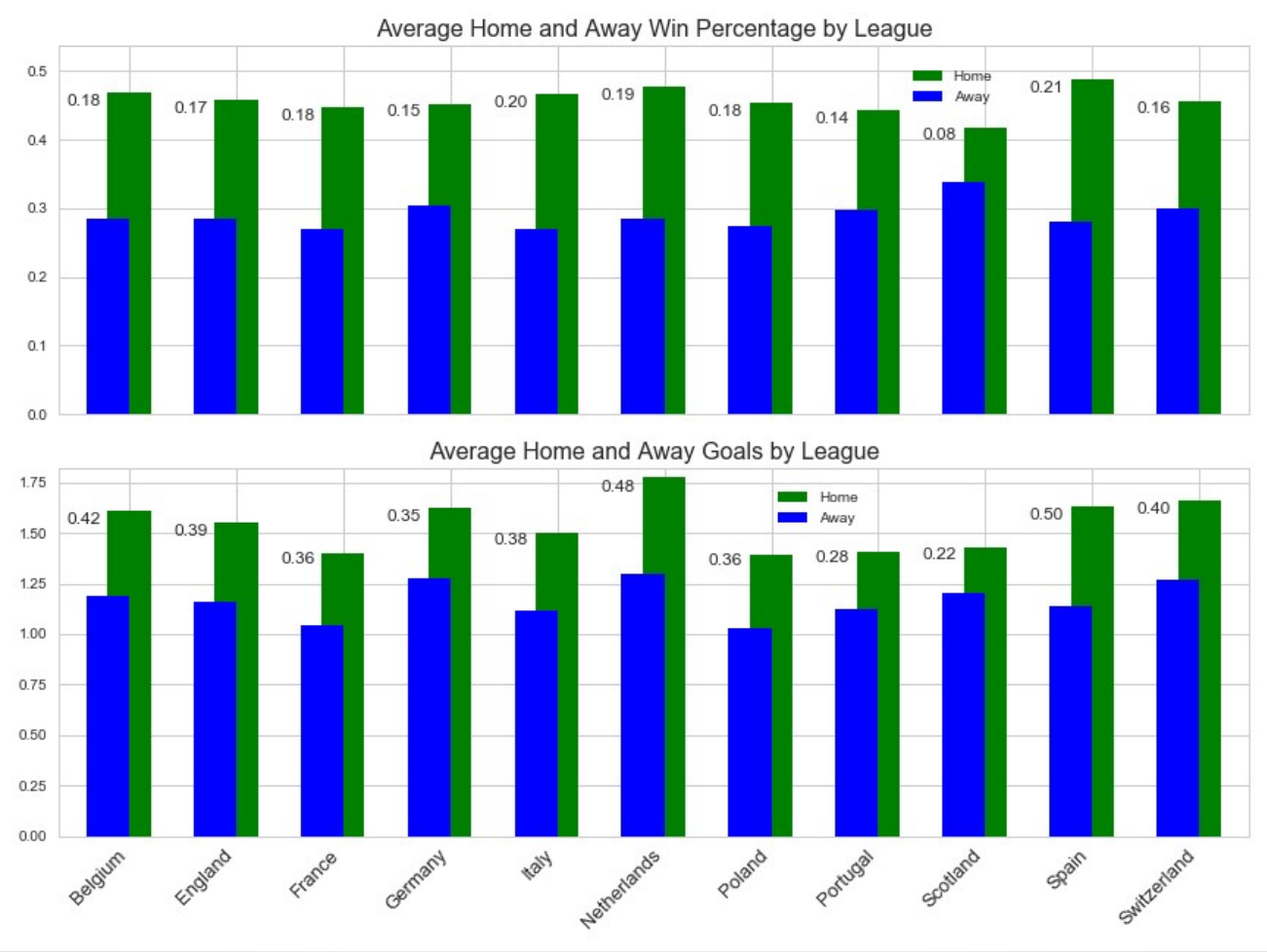
Data Exploration

- Bookmakers' Accuracy
- Home Advantage
- Individual Team Performance
- Goals

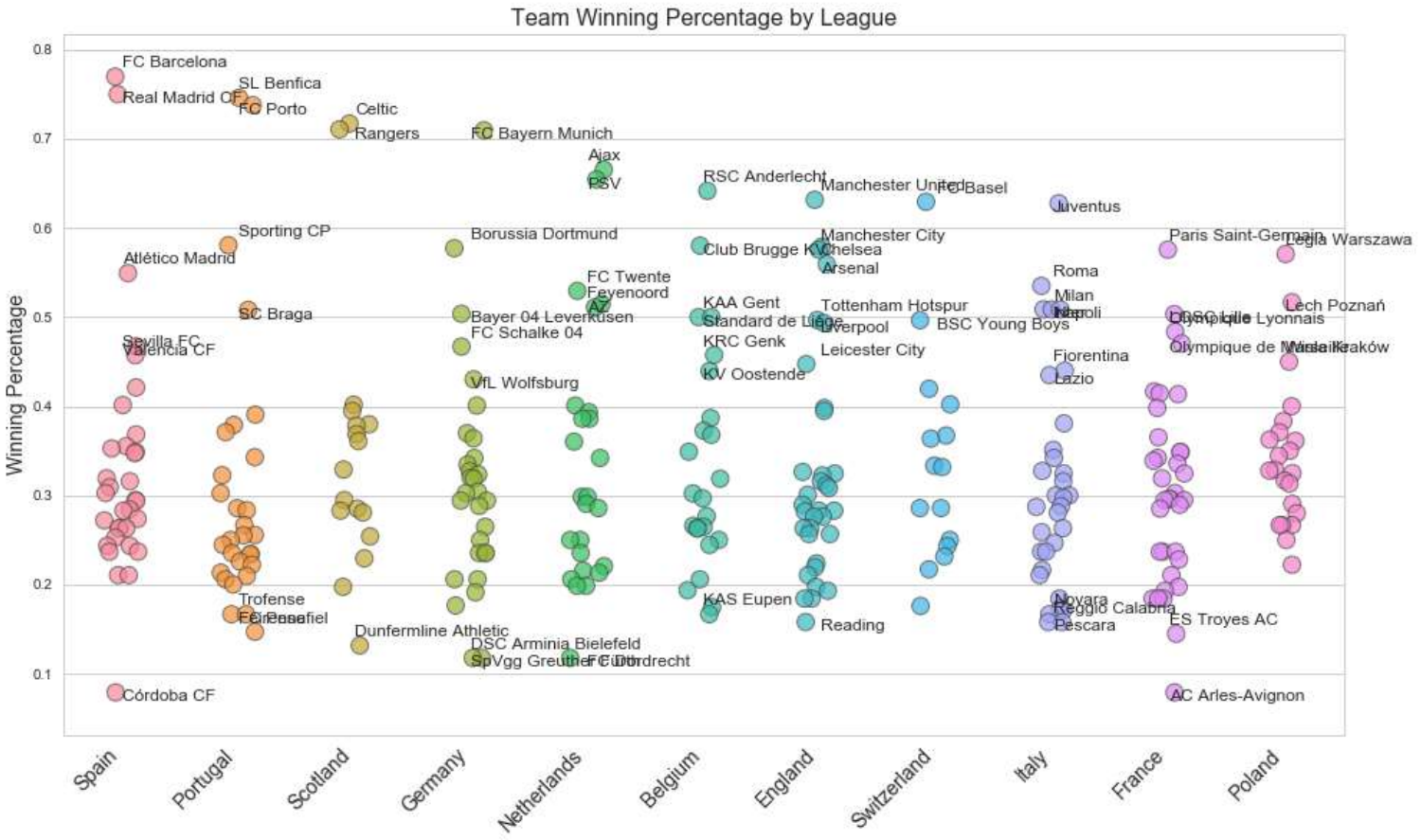
Bookmakers Accuracy



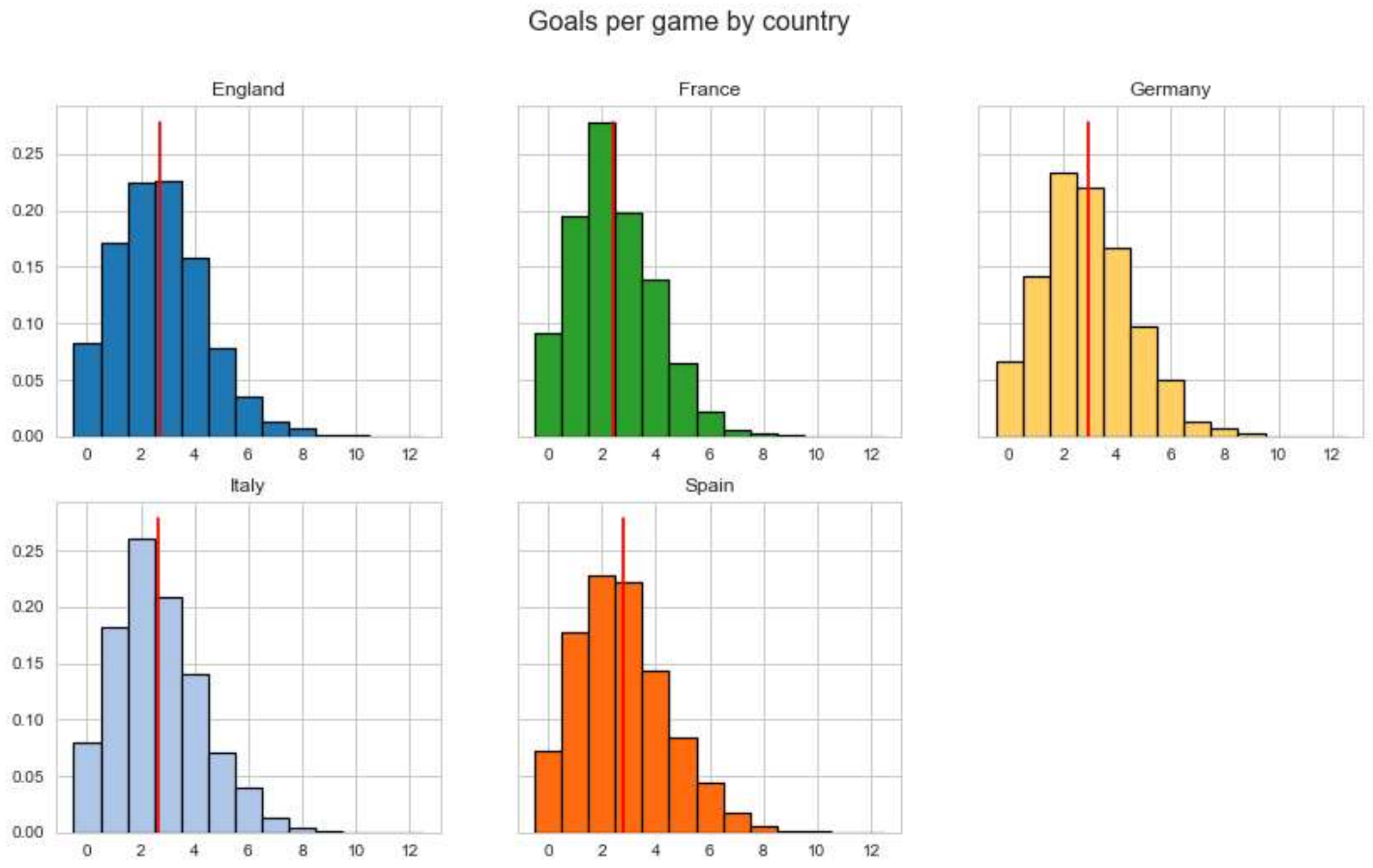
Home Advantage



Team Performance



Goals


Goals per game by country
England
France
Germany
Italy
Spain
0.25
0.20
0.15
0.10
0.05
0.00
0
2
4
6
8
10
12

Features

- Bookmakers's odds: 2 days before game
- Points: average points during current season
- Streak: average points of last 5 games
- Formation: defensive, offensive (binary)
- League and stage: (categorical) and (numerical)
- Overall rating: team average rating from FIFA game
- Best player: rating of best player in team
- Diff: absolute value of difference between max and min odds
- Tie: score on 5 most likely criteria to identify tie games (0 to 5)

Classifiers

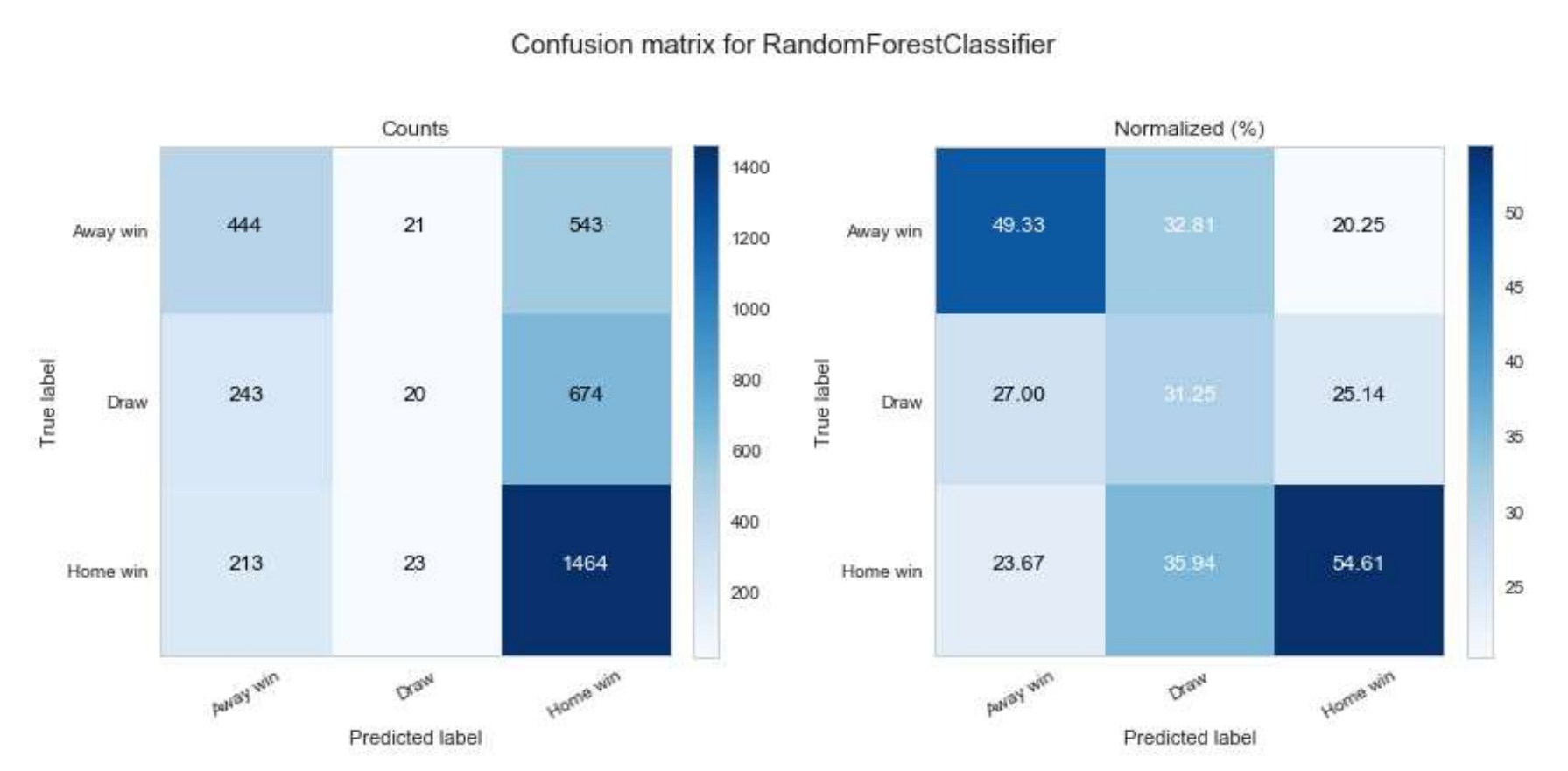
```
LR = LogisticRegression(C=.01, solver='sag', random_state=r,
                        multi_class='multinomial')

RF = RandomForestClassifier(n_estimators=200, max_depth=8,
                min_samples_leaf=6, min_samples_split=20, random_state = r,
                    max_features='sqrt')

ABRF = AdaBoostClassifier(
        base_estimator=RandomForestClassifier(n_estimators=30, max_depth=6,
                min_samples_leaf=3, min_samples_split=10, random_state=r,
                    max_features='sqrt'), n_estimators=15, learning_rate=.35)
```

Log Loss and Accuracy Evaluation

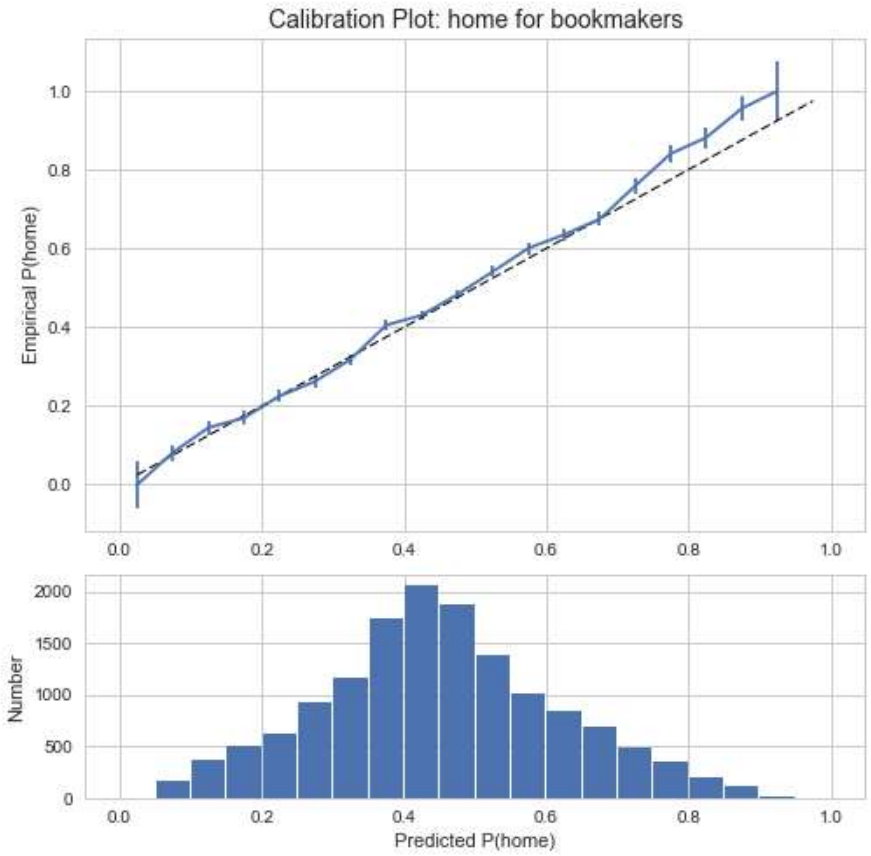
	LogisticRegressionCV	LogisticRegression	RandomForestClassifier	AdaBoostClassifier
Accuracy Test score	52.592593	52.386831	52.784636	52.524005
Accuracy Train score	53.169594	53.229052	55.374131	58.100073
Log Loss Test score	0.984087	0.984294	0.983665	1.046829
Log Loss Train score	0.975790	0.974779	0.910624	1.021158

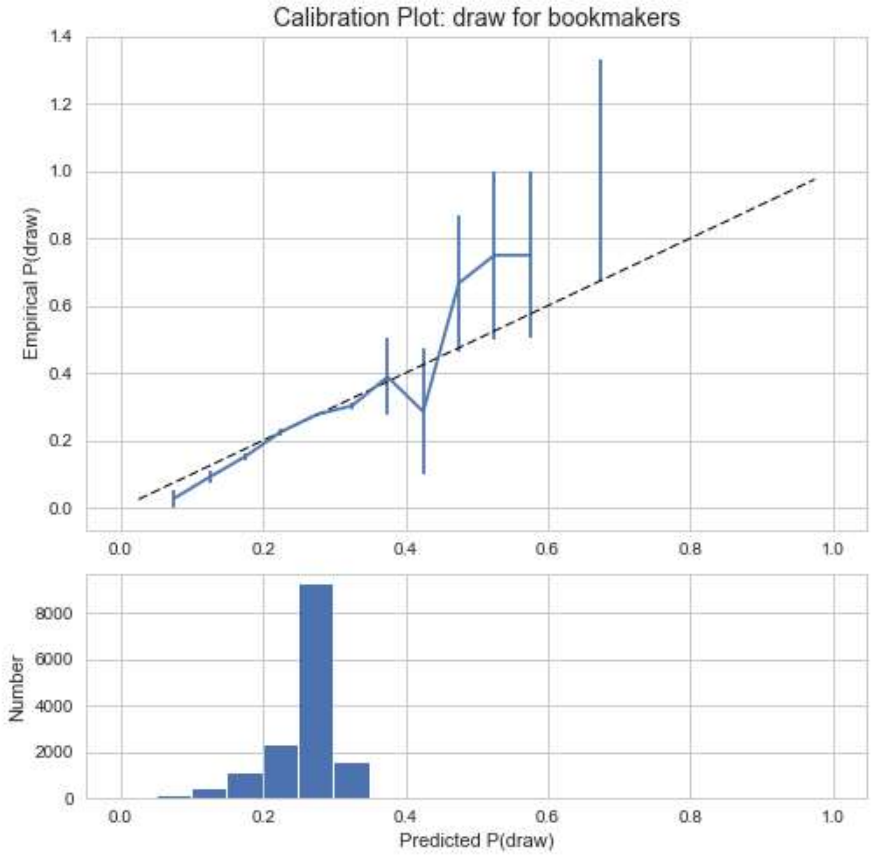
Bookmakers' Accuracy: 52.73 (on test set)

Log Loss and Accuracy Evaluation

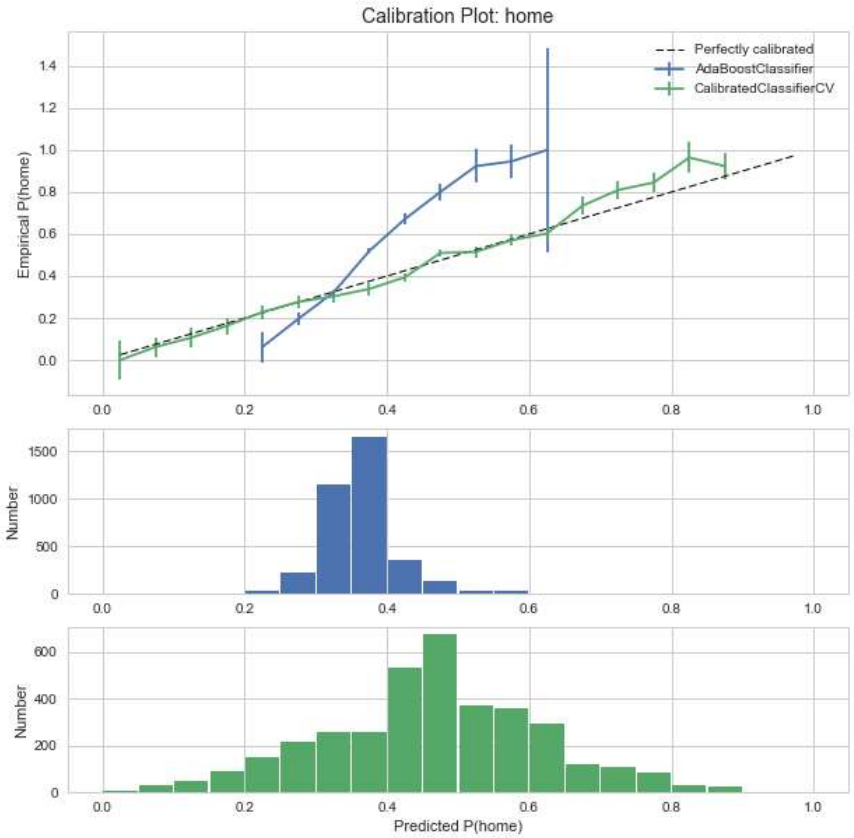


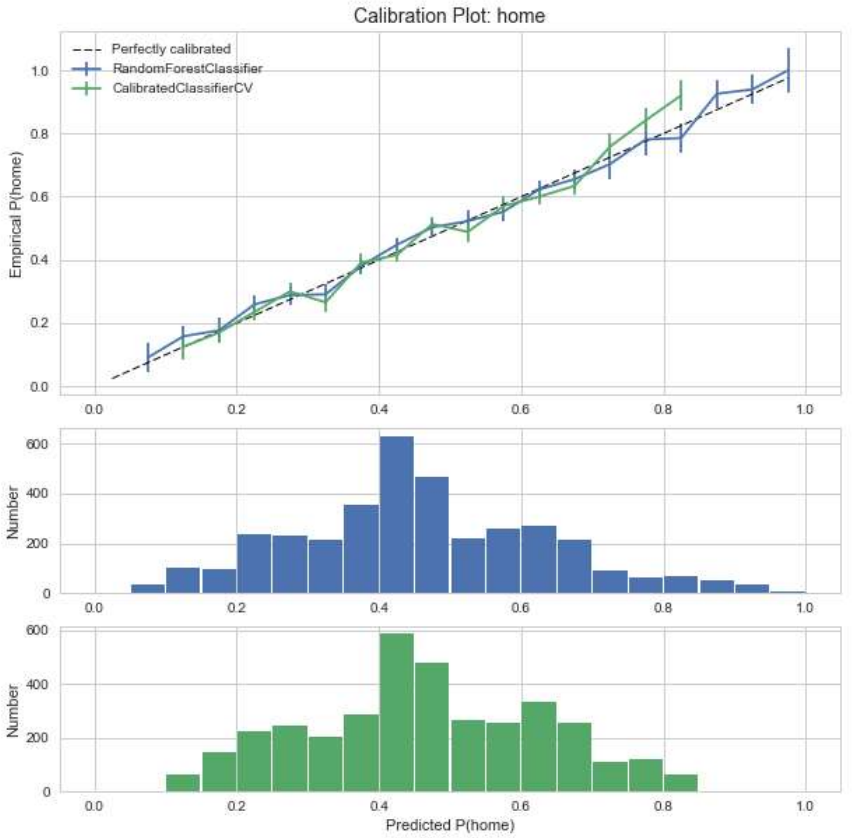
Probability Calibration





Probability Calibration





Return On Investment (ROI)

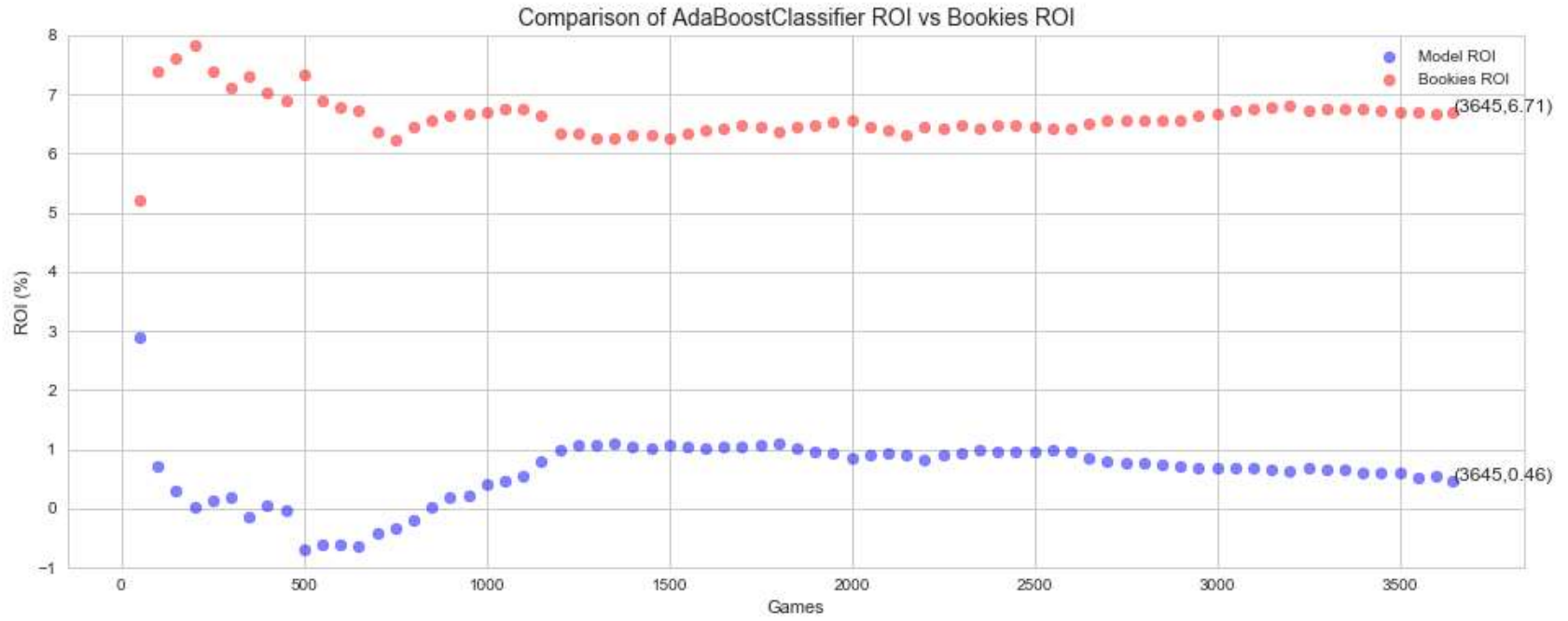
Assumptions

- Bookmakers' and Model impose same commission (6.11%)
- Bet amounts are balanced: same payout regardless of outcome

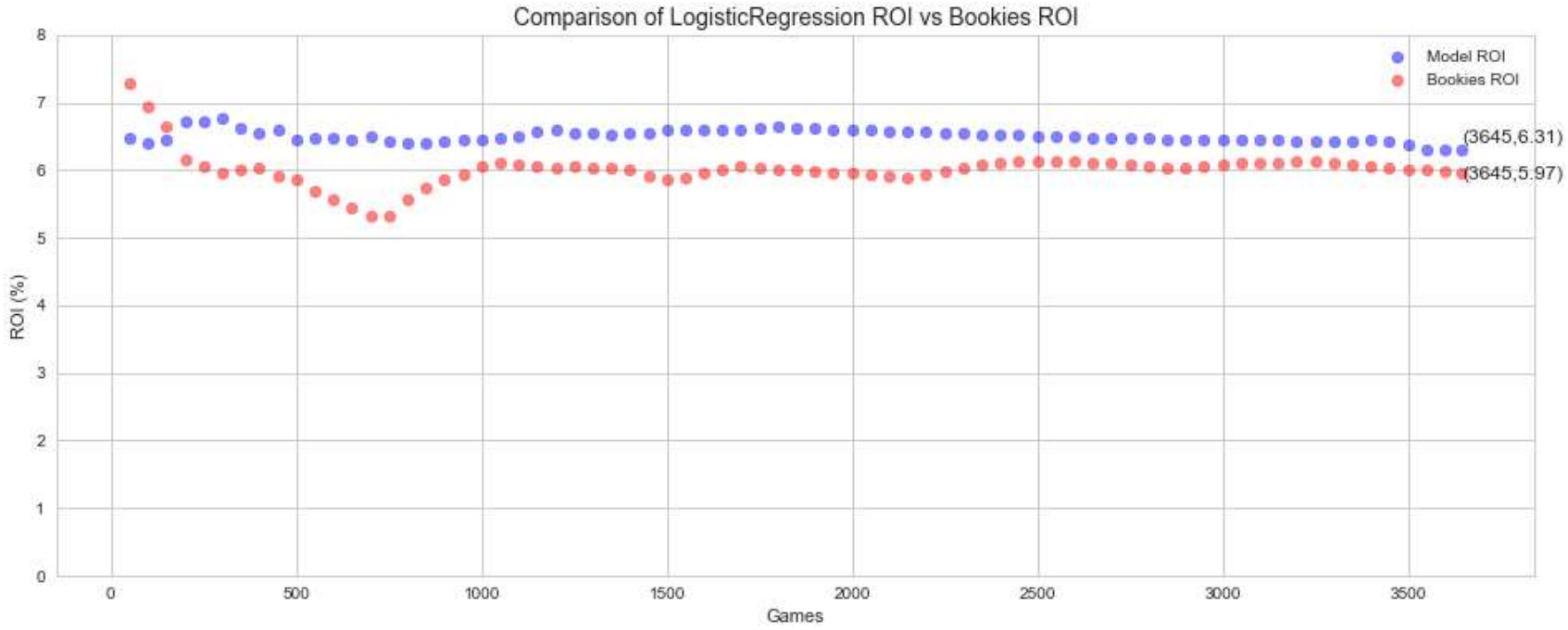
Comparison

- Bet amount is average of balanced amounts for bookmakers and model
- ROI includes any commission

Return On Investment (ROI)
ADABoost (not calibrated) w/ RF



Return On Investment (ROI)
LogisticRegression (not calibrated)



Conclusions

- Best Classifier(s): LinearRegression and/or RandomForest

- Model vs Bookies:
  - Accuracy: 52.78% vs 52.73%
  - ROI: 6.34% vs 6.03%

Recommendations

- Use LogisticRegression or RandomForest, without calibration, and/or boosting.
- Do not use Accuracy score to evaluate probabilities, log loss or ROI are preferred.
- Make sure classifiers are well calibrated on extreme probabilities.
- Determine better features than player attributes from FIFA games. Use real game statistics.