Soccer Match Prediction Models Comparison

Mathias Are, Hans Matthias Andreas
University of Tartu
December 2020

Introduction

In this project we attempted to find out if soccer match results are predictable using machine learning algorithms on data about player individual quality and skill ratings and team synergy characteristics given in the European Soccer database in Kaggle.

The goal of this project was to train two models for predicting the match results: one that uses only the data about individual players and other that uses only the general attributes about the playing teams. After training the models we planned to compare them and see if we can conclude anything about the accuracy of the player and team attributes in EA FIFA video games and also if the given data favors the importance of individual players or team synergy in football.

Methodology

- We used Python as the primary language for data processing.
- The packages we used were pandas, numpy, sklearn and matplotlib.
- Our data preparation process involved first removing unimportant data from our datasets and removing / fixing NaN values. Afterwards we created a dataset for each model by merging the datasets they used and fixed any issues which occurred during the merging.
- Classification algorithms we experimented for predictions were RandomForest, AdaBoost, MLP, NaiveBayes, SVM
- Data analysis involved plotting roc curves, distribution tables, confusion matrices and analysing data correlations to match results

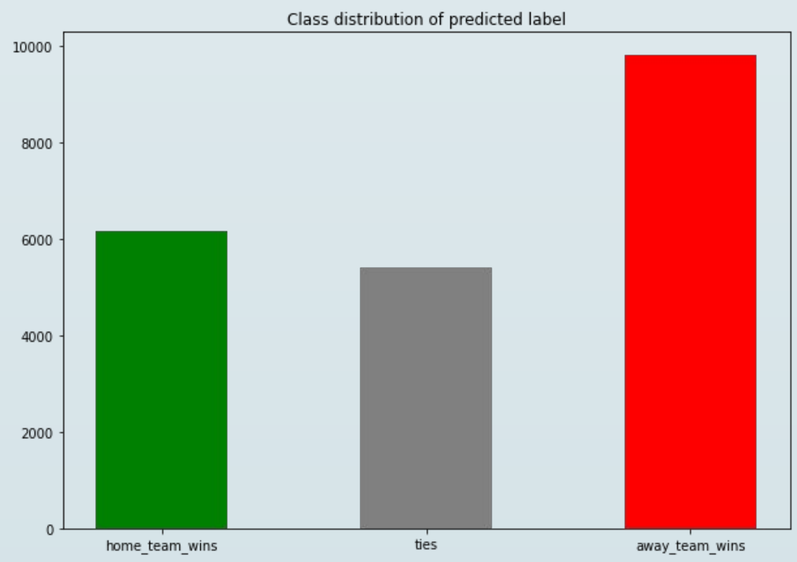
Data

The European Soccer Database from Kaggle contains detailed information about soccer leagues, nations, matches, teams, players, etc. The data has been collected from over 25000 matches played in 2008-2016 and player / team ratings have been scraped from the EA FIFA video game series of the same time period.

Results

With data containing wins, ties and losses, we didn't manage to create models which predicted correctly over 50% of the time, but all of them got close with the model using AdaBoost having the best results.

With data containing only wins and losses, each classification algorithm managed to predict correctly over 65% of the time, with RandomForest having the best results.



Comparison

To our surprise, both models performed quite similarly in terms of accuracy and reacted similarly in the phase of optimization. The only major difference was that model A did not predict tie situations at all and counted most of them to be wins for the home team, whereas model B predicted and also made false predictions evenly for all three classes.

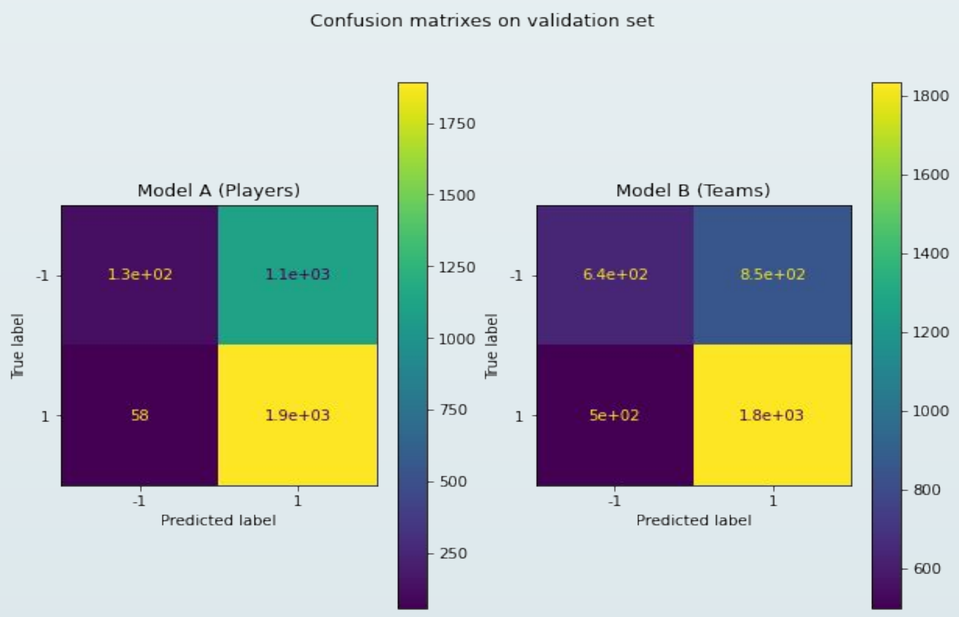
Here is a table about different characteristics of the two models.

Model Name	Accuracy	Precision	Recall	AUC score (ties excluded)	Most successful classifier
Model A (Player attributes)	0.482	0.480	0.482	0.65	Random Forest (sklearn)
Model B (Team attributes)	0.457	0.431	0.457	0.67	Adaboost (sklearn)

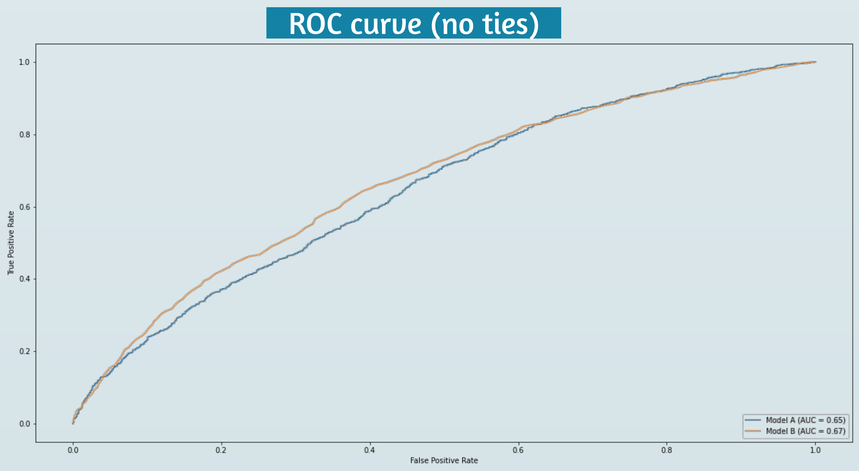
Confusion matrices(ties included)

-1 = away_team_win 0 = tiie 1 = home_team_win

Confusion matrices(ties excluded)



ROC curve (no ties)



Github repository
https://github.com/mathiasare/IdsSoccerPredictionsProject