CS109A Introduction to Data Science

Final Project - 2018 FIFA World Cup

Harvard University

Fall 2018

Instructors: Pavlos Protopapas, Kevin Rader

Team: Group#20 - Nikhil Inamdar, Paul Jureidini, Peeti Sriwongsanguan, Florent Thomas-Morel

Milestone 3 - Revised Project Statement and EDA

1 - Revised Project Statement

1.1 - Objective

Our project objective is to build a model to predict 2018 World Cup match outcomes. The model's accuracy will be measured based on the percentage of outcomes it predicts correctly, and will be compared to a simple baseline model, as well as a model reflecting historical outcomes (e.g. 45% home wins, 30% home losses, 25% home draws).

1.2 - Approach

To accomplish this objective, we will be building three models:

- **Average model:** a model predicting match outcomes based on historical average of home wins, losses and draws.
- **Baseline model:** simple model including FIFA ranking points and a few additional features gathered from the EA Sports FIFA video game. This model will rely on a limited set of modeling approaches.
- **Full model:** a more advanced model based additional on features, and a wider array of modeling techniques.

Basic Model	Full Model	Variables	Description
✔	✔	Match Outcome	Possible match outcomes: Win (0), Draw (1), Loss (2)
✔		FIFA Ranking Difference	Home team ranking points minus away team ranking points (from FIFA.com)
✔	✔	Attack vs Def Rating Difference	Home attack rating minus away rating (from EA Sports)
✔	✔	Midfield vs Midfield Rating Difference	Home midfield rating minus away midfield rating (from EA Sports)
✔	✔	Def vs Attack Rating Difference	Home defense rating minus away defense rating (from EA Sports)
	✔	Overall Record	Exponentially weighted average record vs all teams (from EA Sports)
	✔	Record Versus Opponent	Exponentially weighted average record vs opponent (from EA Sports)
	✔	Home Attack vs Away Def Age Difference	Home attack age minus away age (from EA Sports)
	✔	Home Mid vs Away Mid Age Difference	Home midfield age minus away midfield age (from EA Sports)
	✔	Home Def vs Away Attack Age Difference	Home defense age minus away defense age (from EA Sports)
	✔	Home Attack vs Away Height Difference	Home attack height minus away height (from EA Sports)
	✔	Home Mid vs Away Height Difference	Home midfield height minus away midfield height (from EA Sports)
	✔	Home Def vs Away Height Difference	Home defense height minus away defense height (from EA Sports)
	✔	Home Def vs Away Weight Difference	Home attack weight minus away weight (from EA Sports)
	✔	Home Mid vs Away Weight Difference	Home midfield weight minus away weight rating (from EA Sports)
	✔	Home Def vs Away Weight Difference	Home defense weight minus away weight rating (from EA Sports)
	✔	Attack Chem vs Def Chem Difference	Home attack chemistry minus away chemistry (from EA Sports)
	✔	Mid Chem vs Mid Chem Difference	Home midfield chemistry minus away midfield chemistry (from EA Sports)
	✔	Def Chem vs Attack Chem Difference	Home defense chemistry minus away chemistry rating (from EA Sports. Chemistry is based on how many players play together in club.
	✔	Attack Value vs Def Value Difference	Home attack value minus away value (from EA Sports). Chemistry is based on how many players play together in club.
	✔	Mid Value vs Mid Value Difference	Home midfield value minus away midfield value (from EA Sports. Chemistry is based on how many players play together in club.

Basic Model	Full Model	Variables	Description
	✔	Def Value vs Attack Value Difference	Home defense value minus away value rating (from EA Sports)
	✔	Home Field	Match played on home field (0 | 1)
	✔	Neutral Field	Match played on neutral field (0 | 1)

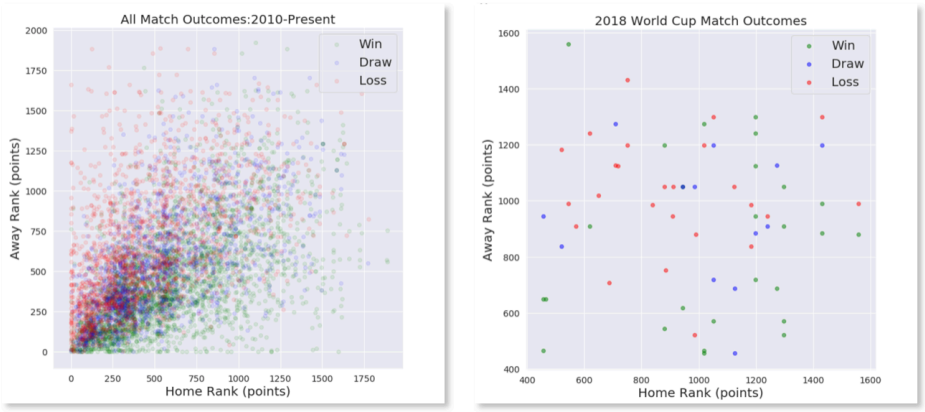
2 - Data

We will be using the following types of data:

- **Known match outcomes**
  - The international match results data set was provided to us, but did not include results from the 2018 World Cup itself. We obtained that data from the FIFA.com website using the Beautifulsoup library and will use this newly acquired data as the basis for our testing dataset.
  - This information will also be used to generate some of the features in our model (Overall Record and Record vs Opponent). These features will be generated using the the pandas.Dataframe.ewm function. We will calculate the weighted average of prior match results (against all opponents and match opponents). The calculation will assign a value of -1 for a loss, 0 for a draw and +1 for a win, but will weigh recent results more heavily.
- **EA Sports ratings**
  - We obtained the EA Sports FIFA game statistics for all national teams and players. The website provides yearly team and player metrics from 2010 to present. This data will be used to obtain features broken out by player position.
- **FIFA rankings**
  - FIFA provides team rankings back to 2010. This data was obtained by scraping the FIFA.com site. Both rankings and ranking points are available. We chose to use ranking points rather than actual rankings in our baseline model because points give a better sense for the distance between the two teams.
- **Other notes**
  - Getting data from different sources meant that team names had to be reconciled prior joining datasets to avoid unintentionally dropping data.
  - We used the pandasql library to join the international results data set with FIFA and EA Sports FIFA data. This allowed us to line up matches results with rankings and ratings as of those dates. This was done by calculating effective and expiration dates for ratings and ratings, and joining datasets where “match date between data effective and expiration date”.
  - We haven’t yet looked into data augmentation techniques. EA Sports ratings are not available for every team, so this limits our training dataset to some extent.

3 - Visualizations

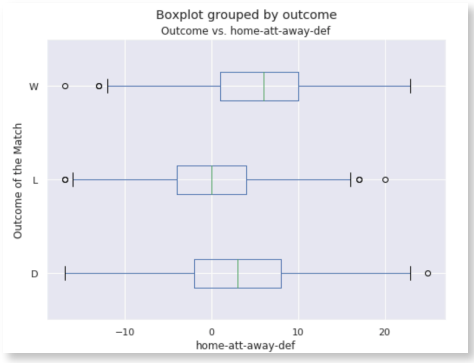
The plot below provides a good sense for the challenge in predicting World Cup match outcomes solely based on FIFA ranking points, and illustrates how more information will be needed to improve the model. There is quite a bit of overlap between the three classes (W,D,L) even though it is apparent that teams with a higher FIFA rank tend to defeat teams ranked lower.

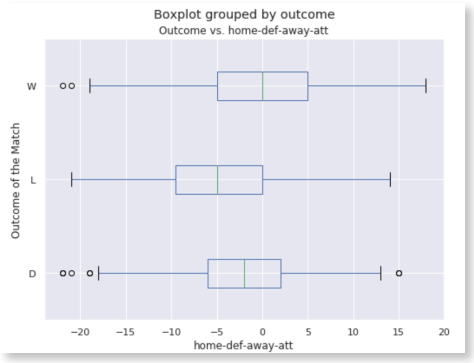


The plot below shows Outcome vs Home Attack vs Away Def Age Difference, and Outcome vs Home Defense vs Away Attack. This illustrates how stronger attack of the home team (compared to the defense of the away team) leads to either the home team winning or drawing the game on average. We can also

see a couple of outliers in the plot. For example, we can clearly see two matches that resulted in win in spite of the home team having poor attack compared to the opposing defense.

On the flip side, if the home team's defense is as good as or even just a little worse compared to the away team's attack, then the home team manages to avoid a loss/defeat. We do see a few outliers in this plot too. There are 2 matches that ended in wins in spite of home defense being very poor.





4 - Initial Baseline Model

We constructed several baseline models to get initial results. The model results can be seen below:

Model Type	Model Parameters	Training Results	Test Results	Confusion Matrix
Logistic Regression	multi_class = 'ovr', cv=5, penalty='l2'	0.506	0.531	Predicted 0 2 All Actual 0 25 1 26 1 11 2 13 2 16 9 25 All 52 12 64
LDA	cv=5	0.503	0.562	Predicted 0 2 All Actual 0 23 3 26 1 10 3 13 2 12 13 25 All 45 19 64
QDA	cv=5	.505	0.594	Predicted 0 1 2 All Actual 0 24 0 2 26 1 10 0 3 13 2 10 1 14 25 All 44 1 19 64