CS109A Milestone 3 EDA and Baseline

Marcus Heijer, Tim Pugh, Will Finigan, Nate Hollenberg

Overview

Our updated project goal is to beat our models that we outline at the end of the document, as well as incorporate these models in an entire World Cup Bracket Simulation; not just predicting independent game results. This will involve using predicted win probabilities. Below is an outline of a lot of the data work that we have done including cleaning and EDA. We understand this is a lot of information, but because of the break we were unable to consistently meet, and wanted to provide as much information as possible for sufficient feedback from you and each other.

Data

Below is an outline of all the data we have collected through different data sources such as publically available data on Kaggle, scraping various websites, and some of the data allotted to us by the teaching staff.

Match and Squad Data

- The first large data set that we gathered were all World Cup Matches from the 2006 World Cup through the 2018 World Cup, called matches.csv. We found a great dataset on Kaggle that had all world cup results from 1930-2014 (the one given to us from the teaching staff had missing games). This was the dataset that we used as the base for our design matrix.
- This dataset was complemented with a squads.csv, which is a dataset full of the rosters for each team competing in a match in the matches data. This dataset has an indicator if the player was a starter and if they scored a goal.
  Cleaning Notes: This data was decently difficult to clean. We got rid of a lot of irrelevant information (useless columns for our model) that was included in the datasets. From the matches.csv we chose our categorization of a result to be Home Team Result, that is if the home team won, drew, or lost the game. Further, we chose to FIFA Country Code (abbreviations created by FIFA), not country name as the main identifier of a team, as country name was very inconsistent among many datasets.

Country Data

- GDP
  We scraped Wikipedia to gather GDP information for every country in the world, with the initial belief that a higher GDP will lead to a better chance of winning. We then used a dictionary we created to put FIFA Country Code in the dataset as well so it would be easy to merge with our other datasets. We got GDP (millions of dollars), GDP (dollars), and GDP per capita for each country as a few of our features.
- Weather
  We scraped Wikipedia to gather average yearly temperature (in Celsius) and average yearly rainfall (in mm) for each country in the world. We then subtracted these numbers from the average temperature or rain of the host country to see how different the climate of the World Cup was from the climate of the nations in the game. We believe that a larger difference in weather might result in a worse performance, as players often perform worse in weather that they are not accustomed to.
- Population
  We scraped Wikipedia to gather population of each country in the world. We believed that nations with a higher population would generally perform better than those countries with smaller populations, because they have more options to choose from for players.
  (All these can be found in gdp_rain_temp.csv)
- FIFA Country Code
  This became the main way that we could identify teams and merge tables as it is the most consistent way to label a team.

Player Data

- FIFA Ratings
  We chose to scrape FIFA player rankings for every player in FIFA 2006, 2010, 2014, and 2018. These FIFA ratings are not incorporated in our baseline model, but we believe will be some of the strongest predictors for our model. Cleaning this data was very hard as well, as we had to deal with strange accents, nicknames, and repeated names (data can be found in 20XX_FIFA_Ratings.csv).

Domestic League Data

- Statistics from all major leagues
  We scraped data from FootballReference to gather player statistics at the club level for the 5 major leagues in England, Spain, France, Italy, and Germany. This data again was hard to clean as we must ensure that the names match up with the player names from the other datasets (squads.csv and FIFA Ratings). We hope to use the player info from these leagues to supplement the player info we have at the international level.

Going Forward

Going forward we plan to use these player statistics and FIFA ratings as very important predictors in our model. Our goal is to get complete FIFA rankings and player stats for each player in each World Cup Match to get a full understanding of the strengths and weaknesses of each team.

External Rankings

- FIFA Rankings
  We scraped the FIFA rankings that FIFA uses to rate National Teams against each other. These were the primary predictor in our baseline model, as our main project goal is to prove to be better than these rankings.
- Betting odds
  We scraped pre-World Cup betting odds for each team. This is a risky predictor to include because Las Vegas has done its work and modeling prior, and betting odds will likely be highly colinear with a lot of our features. We chose to not use betting odds in our baseline model and will carefully select when it is appropriate to use them in our model.

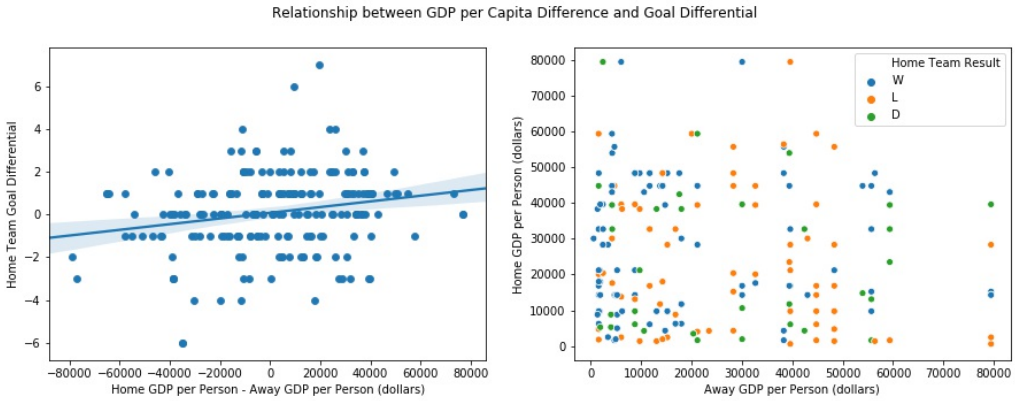
Design Matrix

The design matrix that we created did not include all of the data that we scraped above, as some of it is not complete and we will have to likely perform more data cleaning in the future in order to use such data. We chose to include FIFA Rankings, GDP, weather, and population data in our baseline model.

EDA

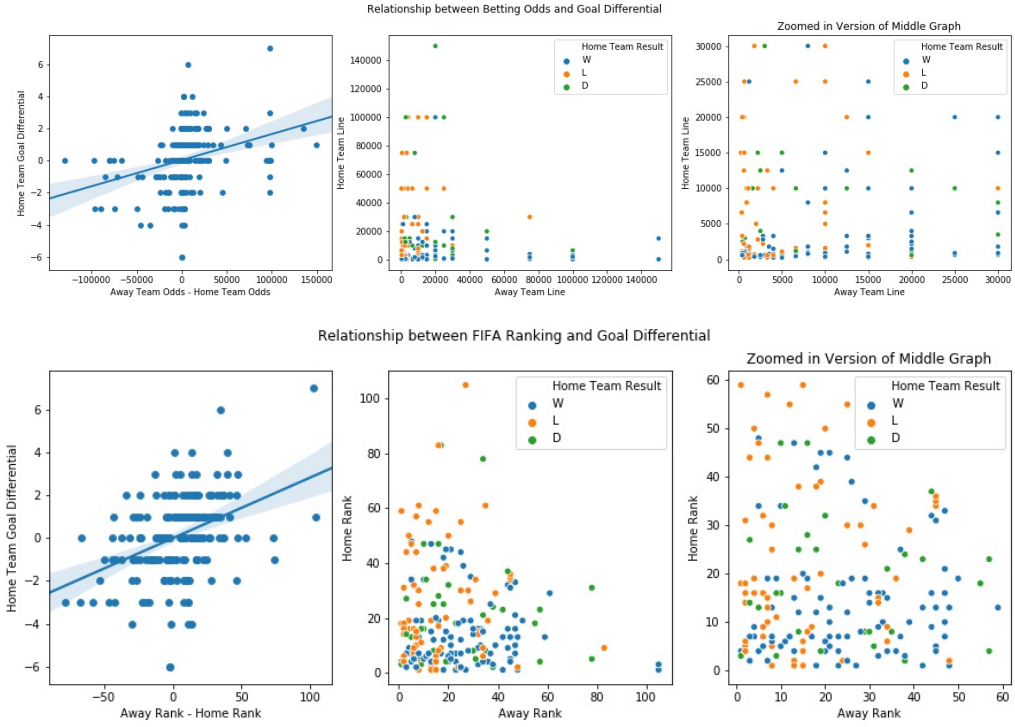
Before making a model, we will first motivate some of our choices for predictors. Specifically, we will focus on the impact of GDP (both total and per capita), population, official FIFA rankings, and pre-tournament betting odds on two response variables: game result (Win, Loss, or Tie) and goal differential.

First, we'll examine the relationship between country GDP per capita and goal differential:



In the first graph, we see a slight positive linear correlation between per capita GDP and goal differential. This follows our expectation since wealthier countries likely have better resources for developing a strong team. In the middle and right plot, we see that the W (Home team win) results are concentrated towards the left side of the plot. Likewise, we see that most of the L (Home team loss) results are located on the bottom of the plot. Both of these results follow our observations that wealthier countries perform better.

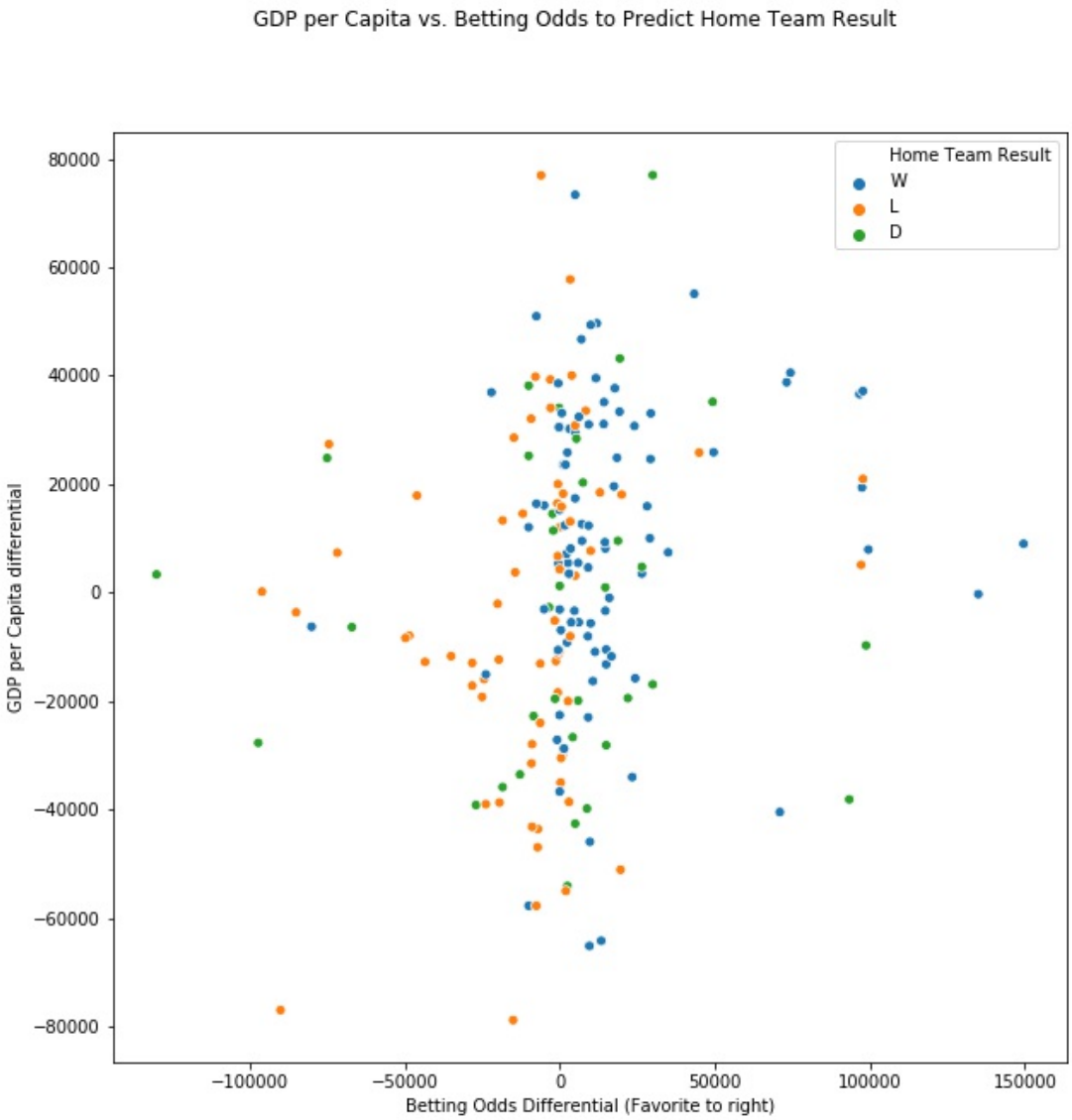
We saw much clearer correlations between Pre-tournament betting odds and FIFA rankings:



Both of the graphs in the left column show that there is a positive linear relationship between odds difference with goal differential and rank difference and goal differential. Home odds and away odds are the pre tournament betting odds. For example, if Argentina's odds are 300 that means if you bet $100 you get $300 if they win. If Panama's odds are 10,000 that means if you bet $100 you get $10,000 if they win. If a team has lower odds that means vegas believes they are more likely to win the tournamnet. The odds difference is the away team odds - the home team odds. That means that the bigger the positive difference in odds the more likely Vegas believes the home team is to win and vice versa. It might be more accurate to do percentage

difference but this will give us a rough gage of how good of a predictor it is. As we can see from the top three graphs it is a very good predictor. There is a pretty strong correlation in the left graph and fairly clear decision boundaries in the right two graphs. But please see the above explanation for why we did not include the Vegas odds in our model.

Looking at the bottom graphs we also see similar things with the relationship between FIFA rankings and Goal Differential. We also see fairly concrete decision boundaries with the scatter plots on the right. This would indicate that FIFA rankings could be a fairly useful predictor in our model.



The graph above shows the betting odds difference on the x-axis and the GDP per capita difference on the y-axis. The X-axis features big home favorites on the right and big home underdogs on the left. We see that Betting Odds is a better predictor than GDP per capita as the position on the y-axis seems to be a minor factor compared to the position on the x-axis. This would make sense as the Vegas odds probably factor in GDP per capita into their odds

along with other predictors. That means the Vegas odds should be better than the GDP per capita.

Baseline models

We created four different baseline models of increasing complexity. We set our prediction goal as accurately predicting each game of the 2018 World Cup independently, at the game level. In other words, we take as input to each game the two competing teams, not taking into consideration the bracket format.

We did not do a whole bracket, as we just predicted at the game level.

- Baseline (Higher Rank Wins)
    - Accuracy Score: 31.3%
- Multi-Class Logistic Regression (OVR): This was the baseline classification model we built. It was multi-class because we wanted to predictor draws in group stage as well.
    - Predictors (for Home and Away Teams): Rank, GDP (dollars), GDP per capita, Temp Difference, Rain Difference, Population
    - Response: Results (W, L, D)
    - Accuracy Score: 43.8%
- KNN regressor predicting goal differential. In this model we subtract Away Team Goals from Home Team Goals prior to training the model, thus making Home Team Goal differential the predictor, and then generating the expected result of the game from the predicted Home Team Goal Differential
    - Predictors (for Home and Away Teams): Rank, GDP (dollars), GDP per capita, Temp Difference, Rain Difference, Population
    - Response: Goal Differential for Home Team
    - Accuracy Score (+ = W, - = L, 0 = D): 48.4%
- KNN regressor predicting home team goals, away team goals, and then seeing who is victorious (or a draw). This was a little complex as we used all predictors to predict the Home Team Goals and Away Team Goals independently, and then subtracted after prediction the generate the expected result of the game.
    - Predictors (for Home and Away Teams): Rank, GDP (dollars), GDP per capita, Temp Difference, Rain Difference, Population
    - Response: Home Goals (for one model), Away Goals (for one model)

- Accuracy Score (after subtracting the two to get result): 48.4%