Exploring Relationships Between Country-Wide Health Markers and World Cup Performance

Cat Johnson
Computer Science
CU Boulder
Boulder, CO USA
johnsocf@colorado.edu

Sabine Hollatz
Computer Science
CU Boulder
Boulder, CO USA
sabine.hollatz@colorado.edu

Greg Giordano
Computer Science
CU Boulder
Boulder, CO USA
grgi1374@colorado.edu

PROBLEM STATEMENT / MOTIVATION

The Féderation Internationale de Football Association (FIFA) has held 21 World Cup tournaments since its inception in 1930. Seventy-nine different countries have qualified to participate in the World Cup in this same timeframe, with eight different countries becoming champions. The purpose of this project is to explore relationships between country-wide markers of physical and emotional health (e.g. life expectancy, positive affect) with performance at the FIFA World Cup (e.g. goals scored, wins, qualifications).

LITERATURE SURVEY (PREVIOUS WORK)

Since 1966, FIFA has conducted a technical study after each world cup. These reports focus on physical skills, psychological attitude, team performance, and on site conditions. Included are descriptive statistics of each athlete and the official results[1].

Goldman Sachs Global Investment Research, as one example of the big investment banks, constructs and trains predictive models for the 2018 World Cup to predict the tournament's outcome. They mined data on team characteristics, individual players, and recent team performances to analyse the number of goals scored in each match[2]. Statista, a market analysis portal, provides statistical analysis for the 2018 world cup with diverse focuses such as average player age, compensation, stadium capacities, and media and fan interest to name only a few[3]. However, neither of those resources considers country-wide markers of physical and emotional health.

Houston and Wilson found positive correlations between country-level income and World Cup appearances, hypothesizing that increased income allows for increased leisure time to increase proficiency in sport[4]. Population was also found to determine World Cup appearances, possibly due to its increasing the likelihood that a star player is born[4].

Hoffman et al. found differences in the economic factors that influence men's vs. women's international soccer performance. In particular, gender inequality as measured by the ratio of women's to men's earnings explained soccer performance[5].

PROPOSED WORK

We plan to approach finding answers to our questions using classic data mining techniques.

These include an important first step of preprocessing our data.

We will account for missing values by replacing them with a measure of central tendency. This step will involve finding consistently the mean or mode depending on the subject, and using this as the replacement value. Using columns of country data, we split up our sets. To determine findings on World Cup qualifiers we will include all countries who play soccer at the international level. To analyze world cup performance we will create a set which includes only countries participating in the World Cup.

To address the topic of noisy data we will use data visualization including scatterplots and boxplots to help identify outliers as stray items. We will trim down our data sets to include only attributes that pertain to questions we will be answering through data analysis.

Merging data sets will be completed by using the country attribute, and the year attribute as items to match on; Allowing us to organize our rows with limitless categories which each uniquely include one country and one year. Since countries may be referenced with slight differences we will create a mapping which includes all referenced names for each country by country, and use that for cases where matching on a country name isn't a direct match. This model sits between multiple data sets and is an interface belonging to the resultant merged dataset.

Dimensionality reduction will be achieved by aggregating wins, goals, and qualificational data when necessary due to an overwhelmingly large dataset. Or to create new rows with which to test trends. We will also use forward stepwise selection as an association rule generating technique to select for sets meeting specifications for strong rules which qualify as interesting based on support and confidence levels and a determined threshold.

To determine correlation we will look carefully at this process of mining single dimensional and multidimensional association rules between data attributes including country related factors and world cup results. We will take this a step beyond finding interesting sets to calculate the Chi-squared (x^2) to determine the statistical significance of our findings, thereby concluding their relevance and quality in finding correlation in data. We will use the Apriori Algorithm to approach finding these sets using a complementary technique to help us mine our the most important sets using the threshold level for pruning out uninteresting correlations. We will keep in mind that correlation doesn't imply causation and work to understand our findings from as centered a source of truth as possible.

To look at some options and exciting possibilities from which to predict results, keeping in mind the stochastic nature of large sporting events, we will look for general trends using Decision Trees and the Bayesian Classifier. These will allow us to group indicators that serve as predictive attributes for successful and qualifying World Cup teams. We will look at how these compare between men and women's soccer, and at what sorts of attribute value pairs create an optimal set, and to which countries these capabilities may belong.

Previous research has examined economic and political factors that are related to performance in the World Cup. While related, this project is unique in that it will specifically look at how measures of country-level health are related to performance.

DATASETS

Measures of World Cup performance will include home goals, away goals, wins, and qualifications, among others. Results from all of the World Cup matches by country and year for all years 1930-2014 will be accessed from Sports Viz Sunday:

https://data.world/sportsvizsunday/sports-viz-sundays-2018/workspace/file?filename=World%20Cup%20Results.xlsx.

Data scraped from sofifa.com on 18,000 individual FIFA players including player rating, club rating, sprint speed, and shot power will be accessed from:
https://data.world/raghav333/fifa-players
And:
https://data.world/sawya/football-world-cup-2018-dataset

Country level health statistics from 258 countries from 1960-2015 including life expectancy, malnutrition, and health expenditure will be accessed from the World Bank:
https://www.kaggle.com/theworldbank/health-nutrition-and-population-statistics

Measures of emotional health from 2005-2016 by country and year including positive affect, negative affect, and social support will be accessed from the World Happiness Report:
https://data.world/laurel/world-happiness-report-data.

EVALUATION METHODS

The analytic quality of our discovered correlations and predictions will be determined by different evaluation methods depending on the used mining method. Metrics, such as accuracy, error rate, and precision, will be computed and used to analyze the quality of our Bayesian Classifier and Decision Trees. Statistical significance needs to be established for our found correlations, important sets and chi-squared test results. Confidence intervals will be calculated and provided. Our evaluation results will be visualized in the form of confusion matrices among other techniques. Comparing our correlation results with the results from predictive modeling will provide further insight into our data.

Along the way, findings in our analysis and evaluation may lead to identification of future change and further interest in exploration.

TOOLS

The data will be preprocessed and analyzed using the Python programming language and relevant statistical libraries in a Jupyter environment. The pandas library will largely be used for preprocessing including cleaning and integrating datasets. Analysis libraries including scikit learn and numpy will be used to conduct statistical tests related to correlation and classification, including Chi-Squared Test, Decision Trees (DT), and Naive Bayes. Matplotlib will be used for data visualization.

Milestones

Preprocessing (by 11/03):

- Missing values
- Noise
- Integration
- Dimensionality Reduction

Find correlations in data (by 11/17):

- Determine Correlation between national markers and success of players and teams in world cup:
    - Find: (a => b)(s, c)
    - Support and Confidence.
- Determine if association is strong.
    - Single dimensional associations, Multi-dimensional association rules.
    - Lift
    - Chi-squared (x^2) and determine statistical significance of findings
    - Apriori Algorithm, find consistent attribute sets that determine success in teams.

Build predictive models (by 12/01):

- Decision Trees
- Bayesian Classifier
- Metrics and Confusion Matrix

REFERENCES

[1] FIFA Technical Study Group (TSG), Technical Reports, https://www.fifa.com/about-fifa/who-we-are/official-documents/development/technical-study-group-reports

[2] The Goldman Sachs Group Inc, 2018 The World Cup and Economics, https://www.goldmansachs.com/insights/pages/world-cup-2018/multimedia/report.pdf

[3] Christina Gough, Soccer - Statistics & Facts, https://www-statista-com.stanford.idm.oclc.org/topics/1595/soccer/

[4] R. G. Houston and D. P. Wilson, Income, leisure and proficiency: an economic study of football performance, https://www.tandfonline.com/doi/pdf/10.1080/13504850210140150?casa_token=7X0eEqYC3n0AAAAA:_FJr1GP4nVoSQNnmGQ26Ak5Kkv4eB_meUrserhl19sURJ7zN6lP6pFetmNNJ3O_jDI7VcZ8GVwYJkw

[5] R. Hoffmann, L C. Ging, V. Matheson, B. Ramasamy, International women's football and gender inequality, https://www.tandfonline.com/doi/full/10.1080/13504850500425774