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

Cat Johnson, Sabine Hollatz, Greg Giordano

Description

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, qualification).

Prior 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. https://www.fifa.com/about-fifa/who-we-are/official-documents/development/technical-study-group-reports

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 tournaments outcome. They mined data on team characteristics, individual players, and recent team performances to analyse the number of goals scored in each match. https://www.goldmansachs.com/insights/pages/world-cup-2018/multimedia/report.pdf

Prior Work

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.
https://www-statista-com.stanford.idm.oclc.org/topics/1595/soccer/

However, neither of those resources considers country-wide markers of physical and emotional health.

Dataset 1: World Cup Results

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

From Sports Viz Sunday found on Data World. https://www.sportsvizsunday.com/

Details: 12 columns, 852 rows in table

history of world cup matches from between 1930 - 2014

Interesting Attributes: year, round, home team, away team, home goals, away goals

Downloaded: Yes, to Cat's and Greg's machine.

Dataset 2: World Happiness Report Data

Url: https://data.world/laurel/world-happiness-report-data

From: the World Happiness Report Website: https://worldhappiness.report compiled by the 'Gallup World Poll'

Details: 2 files, 59 columns 1420 rows. Statistics on a range of properties relating to happiness indicators in countries.

Interesting attributes: country, year, life ladder, gdp per capita, social support, life expectancy at birth, freedom to make life choices, generosity, perceptions of corruption, positive affect, negative affect, confidence in national government, democratic quality, most people can be trusted

Downloaded: Yes, to Cat's and Greg's machine.

Dataset 3: Fifa Players

Url: https://data.world/raghav333/fifa-players

From: data world. scraped from sofifa.com (a site with info on football player ratings)

Details: 92 columns, 17,954 rows. Current Fifa Players, rankings, details and stats on those players

Interesting attributes: player rating, player nationality, potential, preferred foot, wage, skill moves, club rating, sprint speed, shot power, dribbling, heading accuracy, ball control, long passing, vision, standing tackle, sliding tackle, gk_positioning, gk_reflexes

Downloaded: Yes, to Cat's and Greg's machine.

Dataset 4: **Football World Cup 2018 Dataset**

Url: https://data.world/sawya/football-world-cup-2018-dataset

From: data world. Amadou Thione

Details:

Players: 715 rows, 2 columns. | Interesting attributes: player, nationality

Player Score: 11,150 rows,. 15 columns | Interesting attributes: player, goals, assists, rating

Player Stats: 9,133 rows, 9 columns. | Interesting attributes: player, played games, played mins

Teams: 211 rows, 3 columns. | Interesting attributes: team by nationality, points

Player and game data from 2017/ 2018 World Cup. Dataset map for player to nationality.

Downloaded: Yes, to Cat's and Greg's machine.

Proposed Work: Preprocessing

Missing values:
- Missing country attributes
  - Fill with global constant, measure of central tendency (e.g. mean), or most probable value
- Missing World Cup data
  - Include countries not participating in World Cup in analysis of World Cup qualifiers vs. non-qualifiers
  - Drop countries not participating in World Cup in analysis of World Cup performance (goals, wins, etc.)

Noise:
- Use data visualization to help identify outliers

Integration:
- Need to merge data World Cup and country-wide health statistic databases
- Multi-index match by country and year

Dimensionality Reduction
- Can explore relationships on a yearly basis or aggregate wins, goals, qualifications
- Attribute Subset Selection
  - Forward stepwise

Proposed work: Find correlations in data

Determine Correlation.

What national markers are correlated with success of players and teams in the world cup?

Given: min support and min confidence

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

Find Correlations in data continued:

Using the **Apriori Algorithm**, find consistent attribute sets that determine success in teams.

How do national happiness and health indicators in a successful world cup team (or teams that made the semi or quarter finals) stack up in relevance to one another.

What are these attributes belonging to a country that lead to success?

Proposed Work: Build predictive models

Build predictions using classification.

Specifically using: **Decision Trees** and **Bayesian Classifier**

Predict national indicators that lead to successful world cup teams based on sets of indicators.

For instance: {country: x, country_climate_indicator_a: good, country_climate_indicator_b: excellent, players_average_ranking: 8.9, players_average_sprint_speed: 9.3} could be the **optimal set** for winning the World Cup or making it into the matches or qualifications, themselves.

List of Tools

Python

Scikit learn library - Python library that wraps **SciPy**, a library of algorithms. Provides methods ML classification of data.

- We will implement **Chi-Squared Test**, **Decision Trees** (DT), and **Naive Bayes**.

Apriori Algorithm - Python library. We will implement Apriori using support, confidence, and lift.

Pandas - Data structures for flexibility in data modeling. Organize data for analysis.

Numpy - Used for large dimensional arrays. Building block of Scikit learn library.

Matplotlib - Python library providing API for graphing statistical results

Evaluation

Do our correlation data mining results differ than the results from predictive modeling? Does our chi-squared test show significance in our discovered correlations?

Determine the correlations and significance, and importants sets found in our mining and ML. Find to what answers, analysis, and further interest in exploration it leads us.