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 a wide variety of country-wide markers of physical and emotional health with performance at the FIFA World Cup.

Included are:

- Country wide markers from 79 countries that qualified for a FIFA World Cup
- Men’s FIFA World Cup scores since 1930
- Women’s FIFA World Cup scores since 1991
- Player skill levels of nearly 18,000 international FIFA players

Data Preparation Work

- Exploratory Data Analysis
  - visualizing attribute distributions with boxplots and bar graphs
- Data Cleaning and Preprocessing
  - Most missing values were replaced with mean or mode (categorical) of the attribute. Team's nationality in FIFA players dataset were filled with player's nationality
  - Outliers were detected using the interquartile range for each attribute
  - Min/max normalization was conducted on all attributes
  - Country names needed to be matched between datasets.
  - Intermediate datasets were created to track countries that qualified for the World Cup each year and to access total goals and wins for each country each year.
  - Datasets were integrated using country and year attributes
  - Dimensionality reduction was achieved by aggregating wins, goals, and qualificational data, and by forward stepwise selection to find strong and efficient association rules.

Data Preparation Work

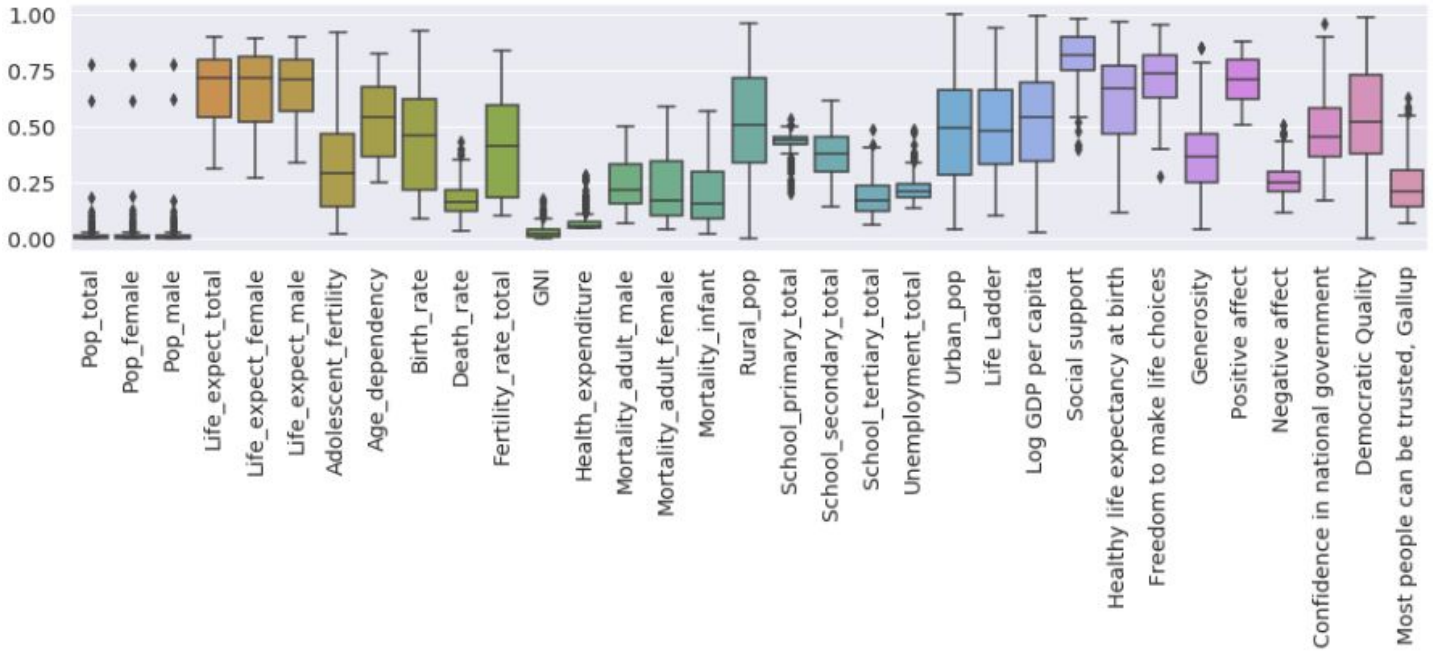


Fig. 1. Country attributes from World Bank data and world happiness report.

Data Preparation Work

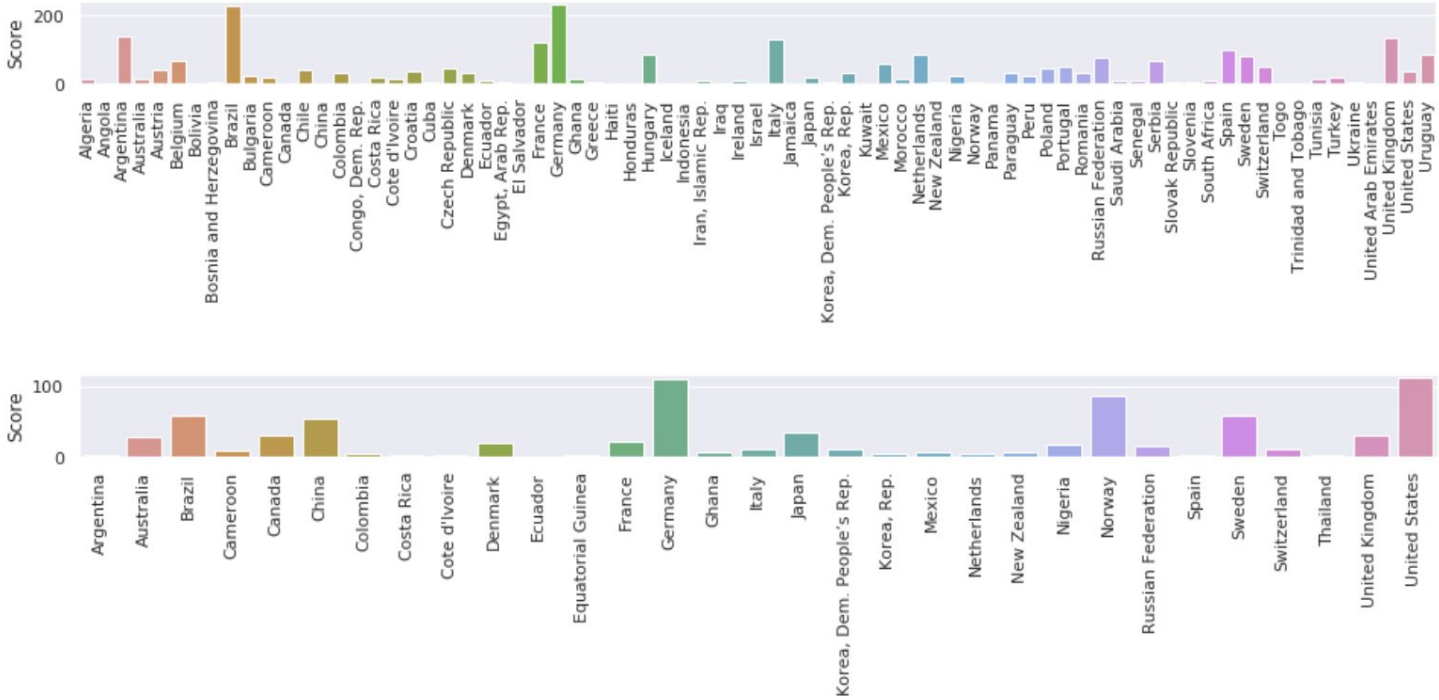


Fig. 3. Comparison of the number of overall goals by country in FIFA men's (top) and women's (bottom) world cups.

Tools Used

Python

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

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 and Seaborn - Python library providing API for graphing statistical results

Key Results - Happiness Report and WC Wins

1. Correlation:
    a. Confidence in National Government.
2. Frequency Sets:
    a. Perceptions of Social Support, Negative Affect (worry, sadness, anger)
3. Prediction:
    a. Best To: 2018 Men's Total World Cup Wins by Country
    b. KNN: 61.5% accuracy at granularity of 95th percentile (20 labels)
    c. Using attributes:
        i. Negative Affect, Most People can be Trusted, Freedom to make Life Choices.

Key Results - National Average Player Skill Strengths and WC Wins

- Correlation:

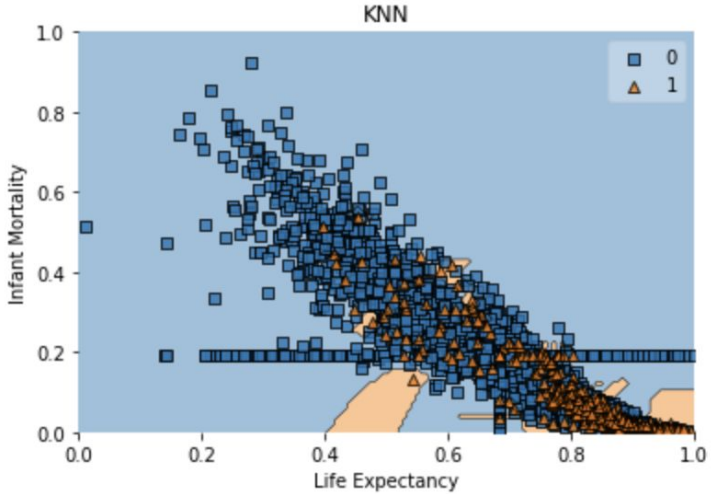
  Positive: Freekick Accuracy, Long Passing, Reactions, Short Passing, Crossing, and Curve.

  Negative: Acceleration, sprint speed, agility, jumping, and stamina.

- Classification:
  - Best to: 2018 Men’s Total World Cup Wins by Country
  - KNN 61% accuracy at 95th percentile, and slightly more successful at predicting lower percentile levels.

Key Results - Classification of World Cup Qualifiers

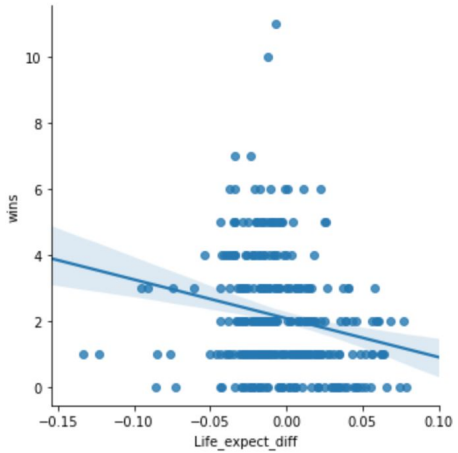
- Imbalanced dataset:
  - ~12% of countries qualify
  - Oversampling
- KNN
  - Accuracy: 83%
  - Precision: 35%
  - Recall: 52%

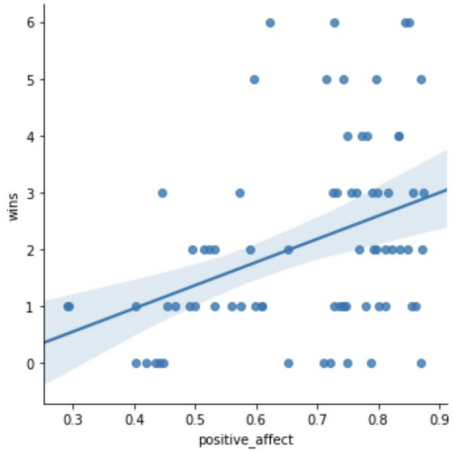



```
[[1161  177]
 [  86   95]]
```

Key Results - Regression

- Generally Low R^2 values (<0.15)
- Significant (p<0.05) values
  - Life expectancy (+)
  - Positive affect (+)
  - Gender difference in life expectancy (-)
  - Gender difference in life expectancy + positive affect (R^2 0.23)





Knowledge Gained and Applications

- Country level markers and performance
- Player skillsets and performance

https://www.google.com/url?sa=i&source=images&cd=&ved=2ahUKEwjsqZSv5bHmAhXQPM0KHU4nC2EQjRx6BAgBEAQ&url=https%3A%2F%2Fwww.fifa.com%2Fworldcup%2Fawards%2F&psig=AOvVaw09jVL0Px_4w5k_aPORpMu0&ust=1576297885054409