Predicting the 2022 W⚽rld Cup Winner

Vania Alvarez Murakami, Suhaib Iqbal, Brian Cahill, Andrew Loose

Team Index: 90

Goal

Use world cup performance data and socioeconomic data to predict the 2022 World Cup winner

Motivation

5,000,000,000

worldwide viewers

227,270,000

viewers per day

Data

Data Collection

Data Needed:

- Matchup Data
  - All Matchups between two Countries in World Cup History
  - Ex: Argentina vs France, etc.
- Socioeconomic Data
  - For each Country and World Cup Year
  - Ex: GDP per Capita, Unemployment Rate, etc.
- Soccer Data
  - Each Country's Performance in Each World Cup
  - Ex: Goals, Assists, etc.

Data Sources:

- FbRef
  - Matchup Data and Soccer Data
- MacroTrends
  - Socioeconomic Data

Data Description

- Each Row Represents a Matchup between Team A and Team B
- For Every Feature, we will have Team A's value, Team B's value, and the Difference of the Two Values
- We will also record whether Team A recorded a Win, Loss, or Draw
- Ex:

Year	Team A	Team B	Team A GDP Growth	Team B GDP Growth	GDP Growth Difference	Result
2018	France	Croatia	1.87 %	2.9 %	-1.03 %	Win

Data Description (Cont'd):

- But wait! We also need to record the matchup between Team B and Team A
- The one row only records Team A's result, not Team B's result
    - We don't want to lose any data when creating our DataFrame
- Ex:

Year	Team A	Team B	Team A GDP Growth	Team B GDP Growth	GDP Growth Difference	Result
2018	France	Croatia	1.87 %	2.9 %	-1.03 %	Win
2018	Croatia	France	2.9 %	1.87 %	1.03 %	Loss

Data Description (Cont'd):

- Overall, we recorded every matchup in World Cup History between a given Team A and Team B
  - Going back to the 1st World Cup in 1930

- For predicting the 2022 World Cup, we have:
  - All matchups in the Group Stage
    - We will build the World Cup Bracket after predicting the winners of the Group Stage
  - Each team's most recently recorded socioeconomic data
  - Each team's soccer stats from 2022 World Cup Qualification

What ML Method Will We Use?

We know the categories we want to predict:

(Win, Loss, or Draw)

Therefore, we will use Classification

Method

- ~~**Decision Tree Classifier**~~
- **K-Nearest Neighbor Classifier:**
  - Wins
  - Losses
  - ~~Draws~~

- **Nearest Neighbors Classifier:**
  - Socioeconomic vs. performance factors
  - All factors

- **Random Forest Classifier:**
  - Feature Importances
- **Nearest Neighbors Classifier:**
  - Top 10 factors

Method

- ~~**Decision Tree Classifier**~~
- **K-Nearest Neighbor Classifier:**
  - Wins
  - Losses
  - Draws

- **Nearest Neighbors Classifier:**
  - Socioeconomic vs. performance factors
  - All factors

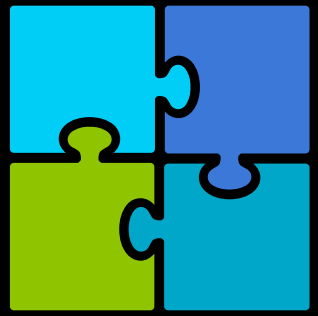
- **Random Forest Classifier:**
  - Feature Importances
- **Nearest Neighbors Classifier:**
  - Top 10 factors

Method

- ~~**Decision Tree Classifier**~~
- **K-Nearest Neighbor Classifier:**
  - Wins
  - Losses
  - Draws

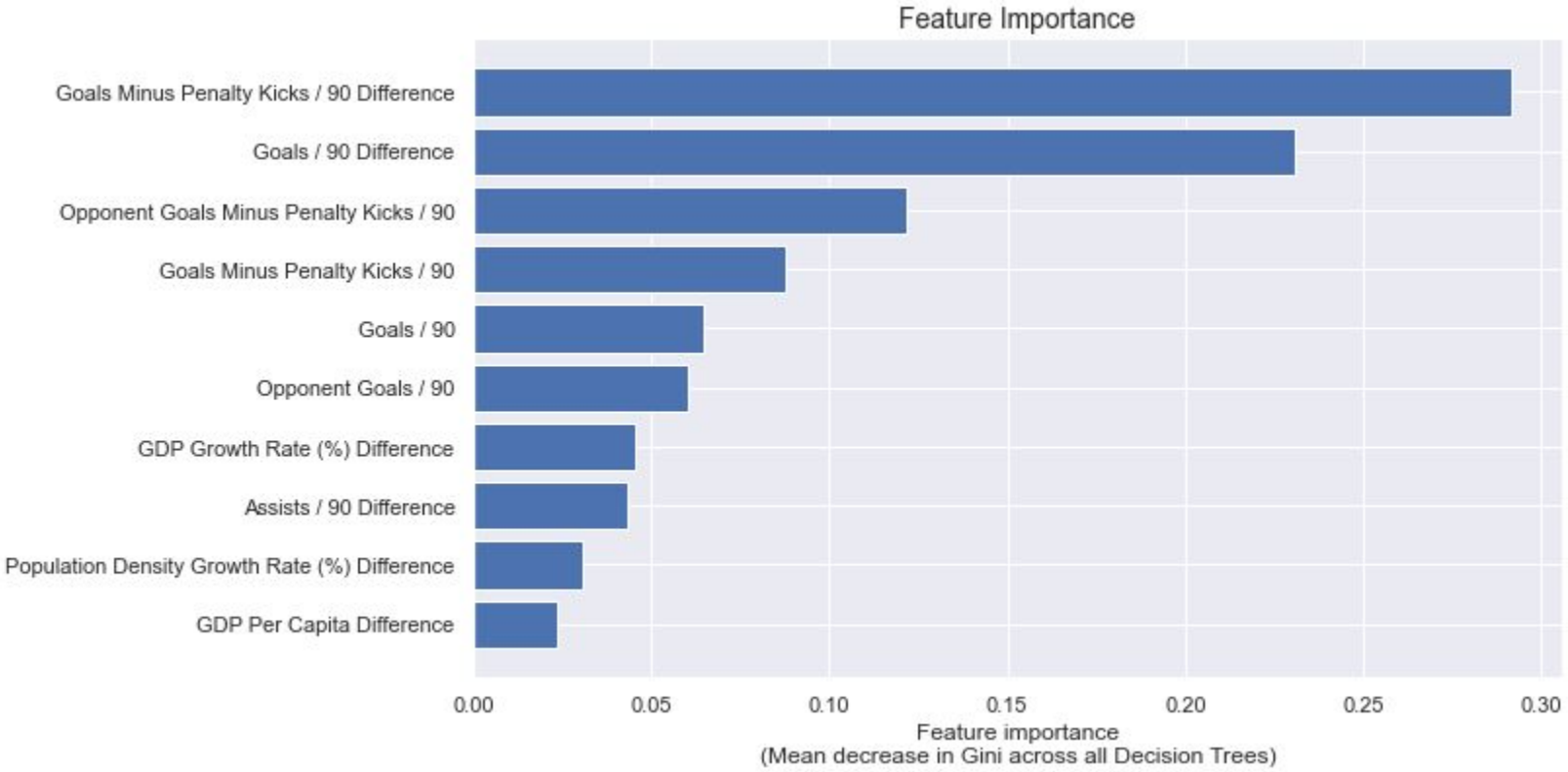
- **Nearest Neighbors Classifier:**
  - Socioeconomic vs. performance factors
  - All factors

- **Random Forest Classifier:**
  - Feature Importances
- **Nearest Neighbors Classifier:**
  - Top 10 factors



Results

Random Forest Classification & Feature Importance

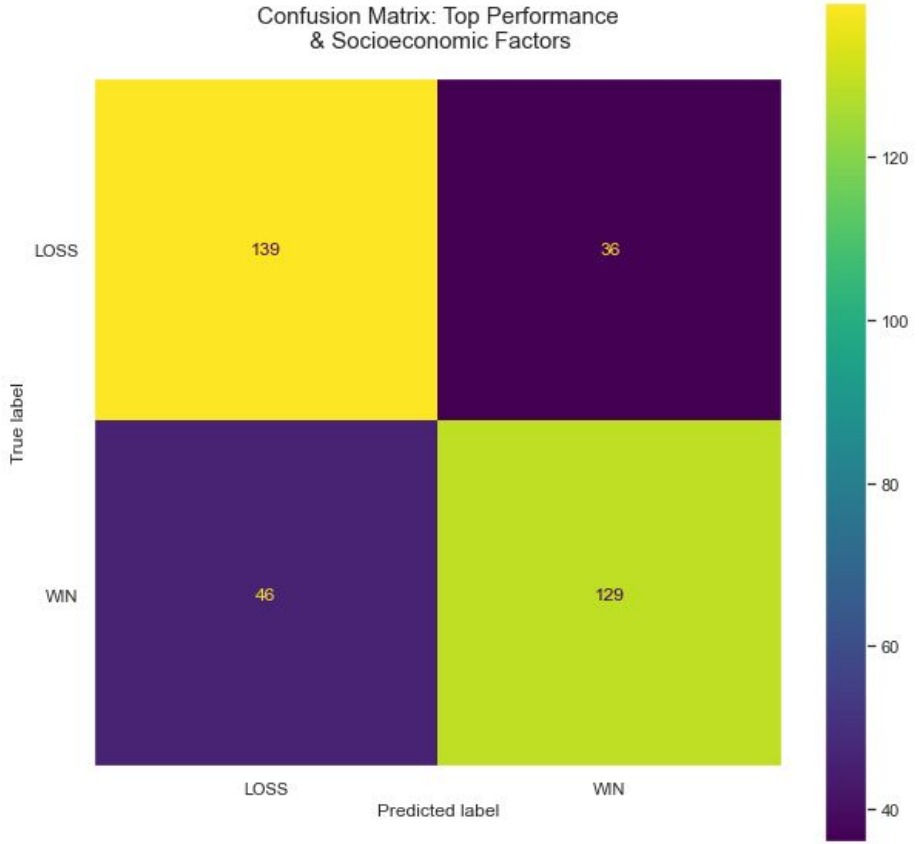


Machine Learning & KNN Classification

Accuracy Score: 0.76

Other Models

- Top Performance Factors
  - Accuracy score = 0.71
- Top Socioeconomic Factors
  - Accuracy score = 0.58



Note. Draws were excluded from model

Our Prediction: 2022 WCF Group Stage Winners

Group A: Netherlands, Qatar

Group B: England, IR Iran

Group C: Argentina, Saudi Arabia

Group D: Denmark, France

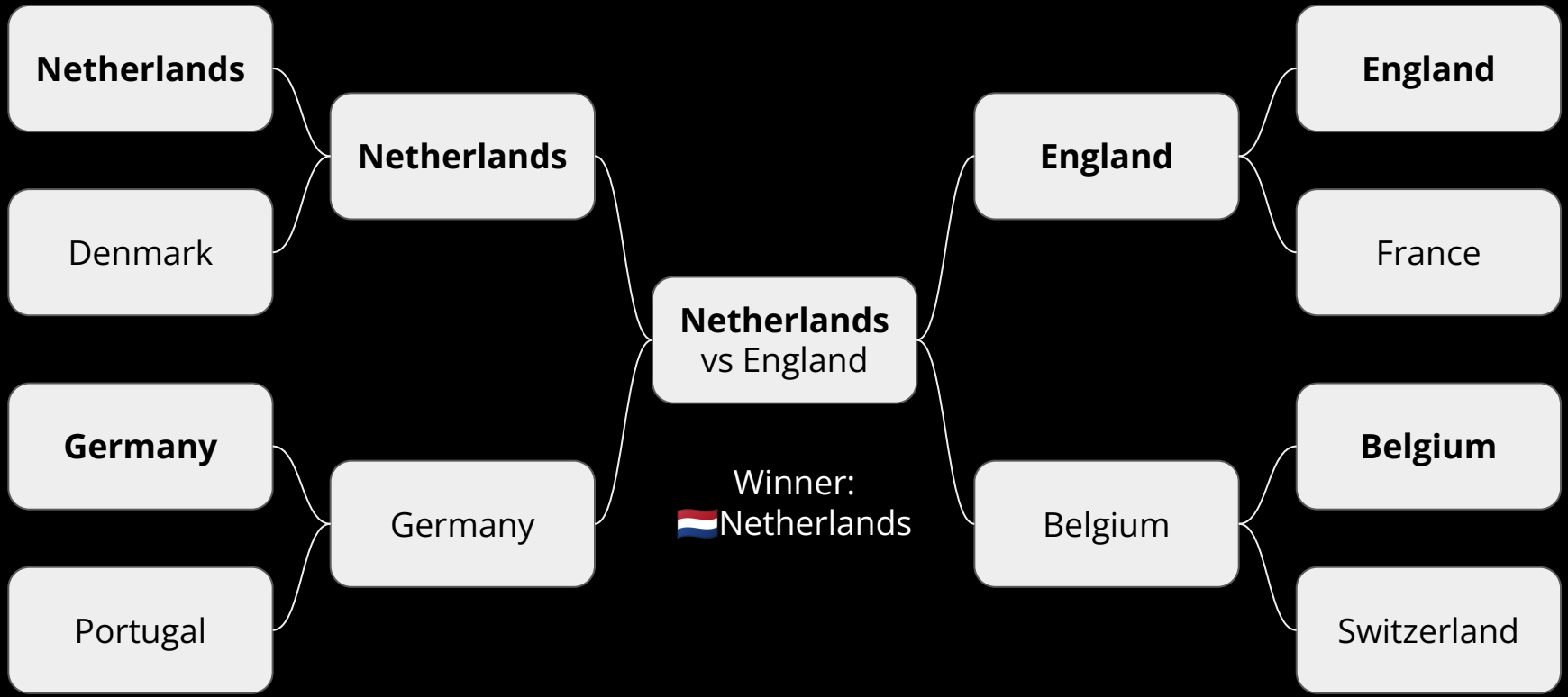
Group E: Germany, Japan

Group F: Belgium, Canada

Group G: Brazil, Switzerland

Group H: Portugal, Korea Republic

Our Prediction



Key Takeaways

Challenges

- Our initial model had extreme difficulty predicting draws
- Much more data available in 2018 that would improve performance of our model

Future Considerations

- KNN classifier built off the 10 most important factors but those factors were all given the same weight
- When training our KNN classifier we only used historical world cup data

Questions?