Technische Universität Berlin

Predicting World Cup 2018 Matches

Tanvir Ahmed

Md Kamrul Hasan

10th July, 2019

Agenda

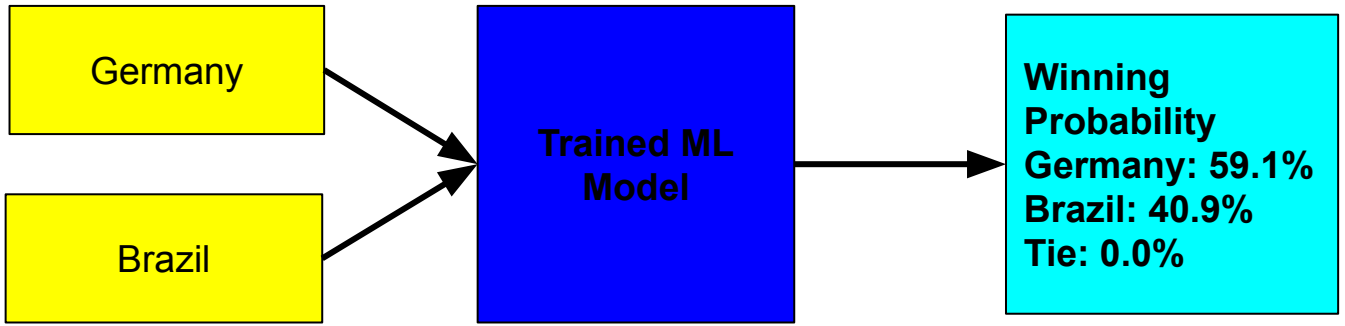
- Objective
- Problem Statement
- System Architecture
- Data Acquisition and Integration
- Data preprocessing
- Model Evaluation
- References
- Demo

Objectives

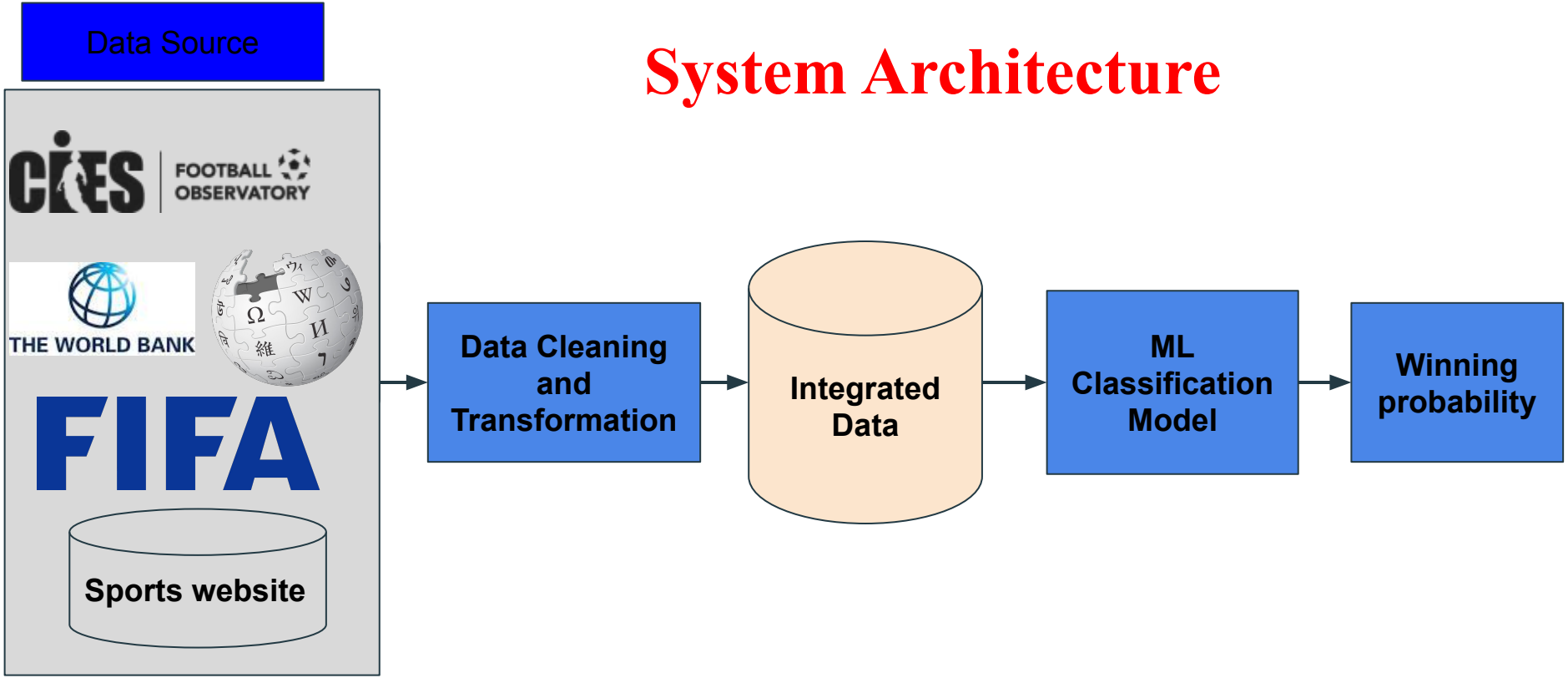
Predicting the winning probability of world cup 2018 matches through collecting data from different heterogeneous sources

Problem Statement

- Collecting featured data which impact on the matches
- Training data using ML Algorithm
- Two teams as input
- Showing their winning probability



System Architecture



Data Acquisition

- **Economic Factors**
  1. GDP per capita 2017(Wiki)
  2. Population 2017 (World Bank)
  3. Happiness rank (World Happiness Report)
  4. Happiness score (World Happiness Report)
  5. Life expectancy (World Happiness Report)
  6. Freedom (World Happiness Report)
  7. Generosity (World Happiness Report)
  8. Government Corruption (World Happiness Report)

Data Acquisition

- **Supportive Factors**
  9. FIFA Rank (FIFA)
  10. ELO Rating (Elo website)
  11. Final appearance (Wiki)
  12. Semi Final Appearance (Wiki)
  13. Last best performance (Wiki)
- **Home Advantage**
  14. Home country advantage (Wiki)
  15. Home continent advantage (Wiki)
- **Team Structure**
  16. Players average appearance (Wiki)
  17. Players average goals (Wiki)

Technische Universität Berlin

FIFA

http://eloratings.net

Data Acquisition

- **Team Structure**
  18. Percentage of players in foreign club (CIES)
  19. Average age of players (CIES)
  20. Soccer power index (FiveThirtyEight)
  21. Star Players (EA sports)
  22. Coach ranking (Real Sport)
  23. Attack side (FiveThirtyEight)
  24. Defensive side(FIFA Index)
  25. Middle side (FIFA Index)
  26. Defensive rating (FiveThirtyEight)
  27. Offensive rating (FiveThirtyEight)
  28. Over all side (FIFA Index)

Integrated Schema

['country','fifa_rank','elo','avg_age','home_country_adv','home_continent_adv','last_best_performance','star_count','coach_performance', 'final', 'semi','foreign_club','att','def','mid','ovr','power_index','offensive','defensive','avg_players_appearance','avg_players_goal','country_happiness_rank','gdp','population','happiness_score','life_expectancy','freedom','generosity', 'goverment_corruption']

Data Preprocessing

Match Result
64 matches

Team-1	Team-2	Team-1 Score	Team-2 Score
Russia	Saudi	5	0
Egypt	Uruguay	0	1
Iran	Morocco	1	0

Features
28 Features

Team	fifa_rank	gdp	...	govt_corruption
Russia	70	60	...	0.03
……..	……..	….	…	...
Belgium	3	18		0.25

Team-1	Team-2	Team-1 Score	Team-2 Score	Team_1 fifa_rank_1	Team_2 fifa_rank_2	fifa_rank	---	winner
Russia	Saudi	5	0	70	67	17	---	1
Egypt	Uruguay	0	1	45	14	31	---	-1

Data Preprocessing

With Duplicate value

Technische Universität Berlin

Target Class

Team-1	Team-2	Team-1 score	Team-2 Score	Team_1_Fifa_rank	Team_2_Fifa_rank	fifa_rank	…	Winner
Russia	Saudi	5	0	70	67	17	...	1
Egypt	Uruguay	0	1	45	14	31	…	-1
…	...	...	...	...	...	...	...	
Saudi	Russia	0	5	67	70	-17	…	-1
Uruguay	Egypt	1	0	14	45	-31	...	1

Training and Testing Data

- Training Data
  - Group stage (48 matches)
  - With duplicates (96 rows)
- Testing Data
  - Remaining stage (16 matches)
  - With duplicates (32 rows)

Normalisation and Scaling

- To reduce the highly varying values in features
- To make all features same level of magnitudes

Three Scaling methods

- StandardScaler
- MinMaxScaler
- RobustScaler

StandardScaler

- Data is normally distributed within each feature
- The distribution is now centred around 0, with a standard deviation of 1

Formula for any feature x:

$$\frac{x_i - mean(x)}{stdev(x)}$$

Formula: [19]

Tiny Features vs Mega Features
Image: [12]

Normalisation and Scaling

MinMaxScaler

- The values are between 0 and 1 (or -1 to 1 if there are negative values)

  Formula for any feature x:

$$\frac{x_i - min(x)}{max(x) - min(x)}$$

RobustScaler

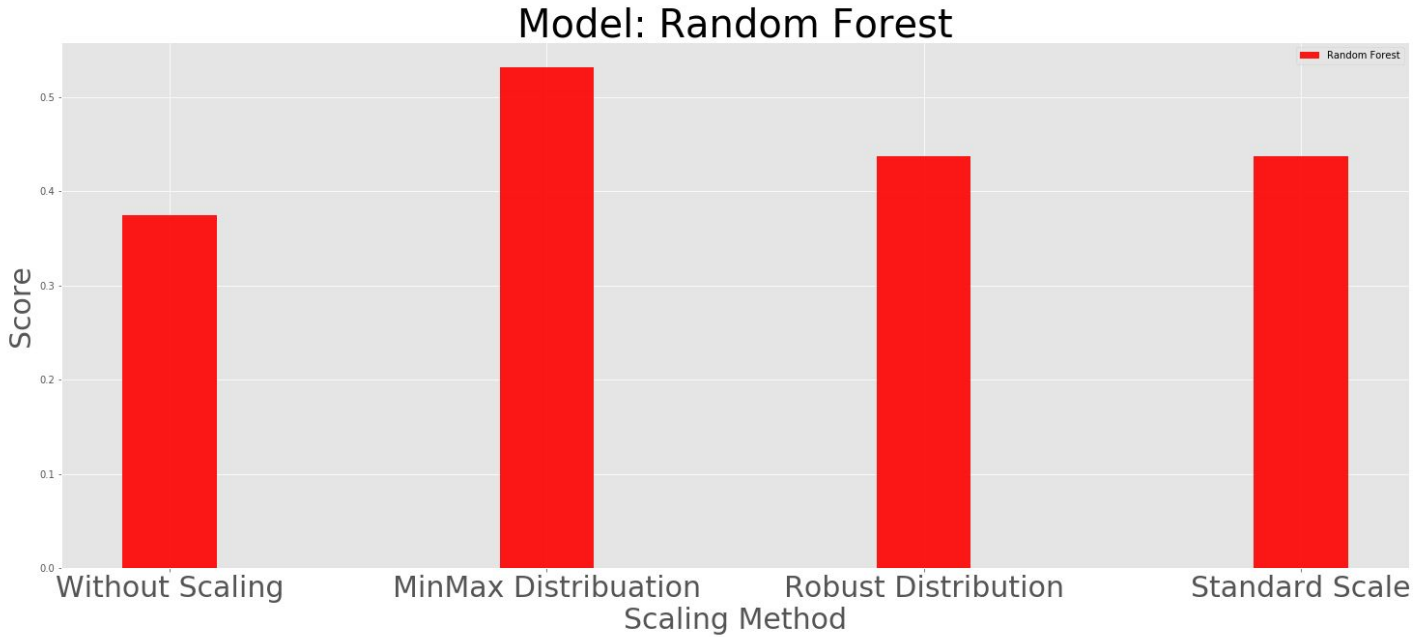
- Similar method to the Min-Max scaler but it uses the interquartile range

  Formula for any feature x:

$$\frac{x_i - Q_1(x)}{Q_3(x) - Q_1(x)}$$

Formula: [19]

Normalisation and Scaling



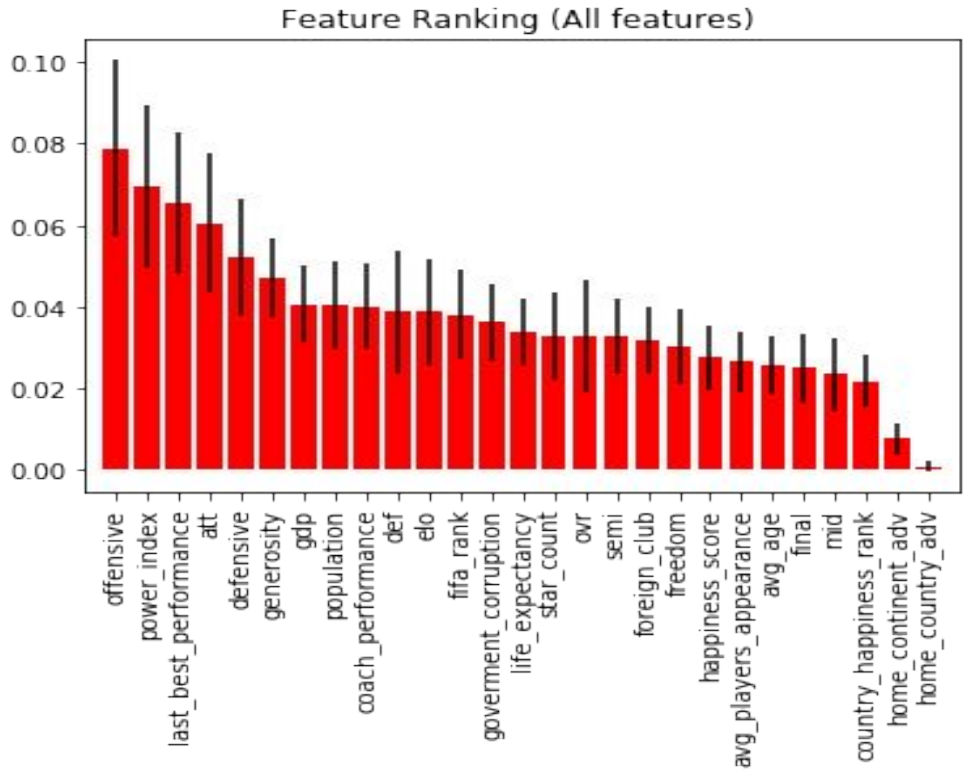
Feature Selection Technique

- Feature Ranking - All features
- Feature Ranking - Individual feature
- Recursive Feature Elimination
- Forward Selection

Feature Selection Technique

Feature Ranking (All)

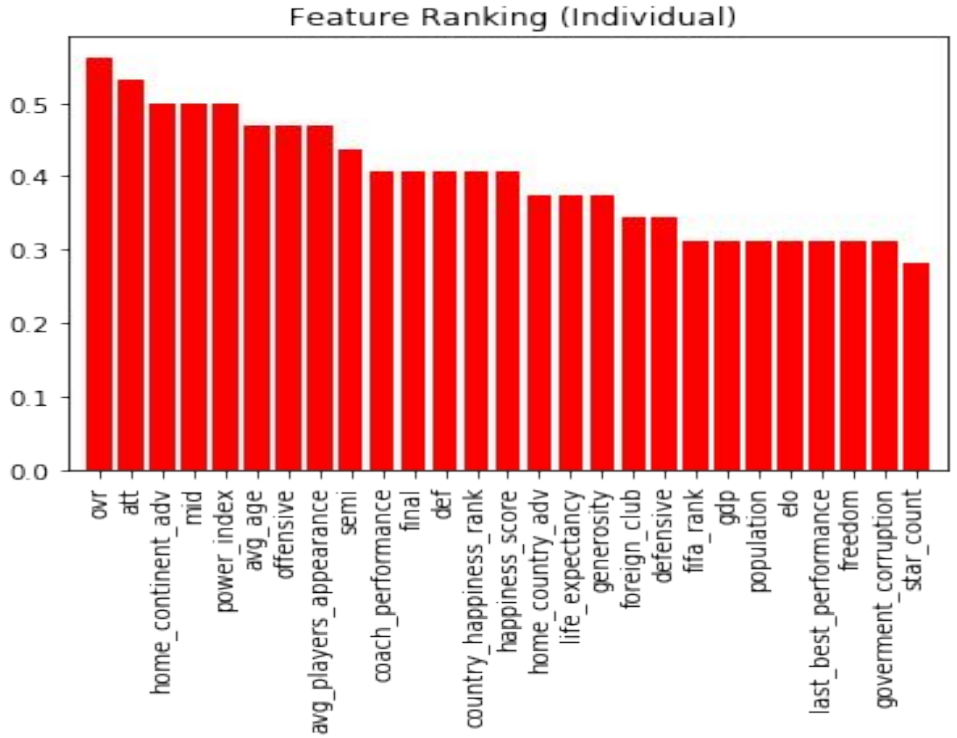
- Model : RandomForest
- Top : Defence
- Base : Home country advantage



Feature Selection Technique

Feature Ranking (Individual)

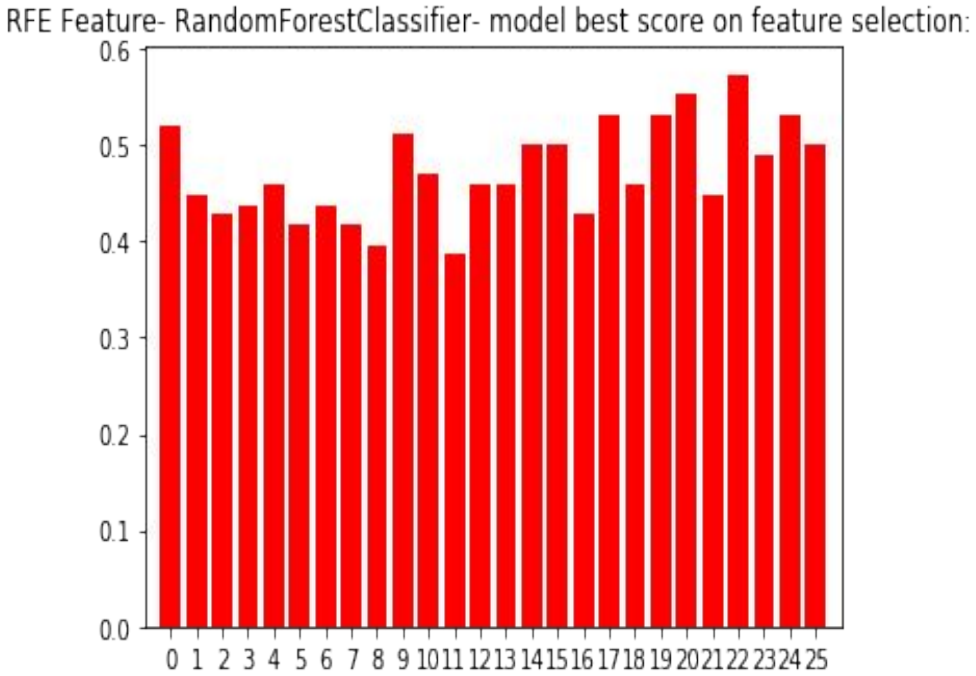
- Model : RandomForest
- Top : Overall rating
- Base : Star player count



Feature Selection Technique

Recursive Feature Elimination

- Model : RandomForest
- Optimum number of features : 23



Feature Selection Technique

Technische Universität Berlin

Recursive Feature Elimination

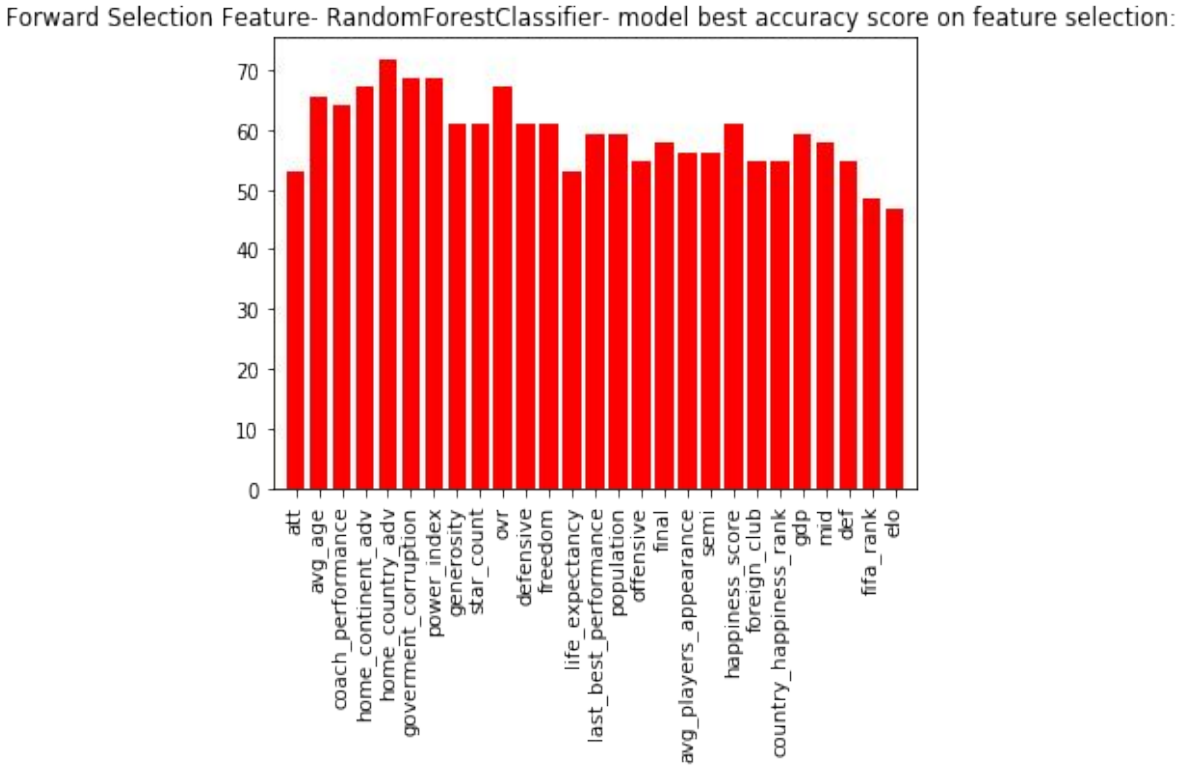
[offensive, power_index, last_best_performance, att, defensive, generosity, gdp, population, coach_performance, def, elo, fifa_rank, goverment_corruption, life_expectancy, star_count, semi, foreign_club, freedom, happiness_score, avg_players_appearance, avg_age, final, mid]

Feature Selection Technique

Technische Universität Berlin

Forward Selection

- Model : RandomForest
- Top : Attack
- Base : ELO rating



Feature Selection Technique

Technische Universität Berlin

Final Features

['Att','avg_age','coach_performance','home_continent_adv','home_country_adv','goverment_corruption','power_index','generosity','star_count','ovr','defensive','freedom',’life_expectancy','last_best_performance','population','offensive','final','avg_players_appearance','semi','happiness_score','foreign_club','country_happiness_rank','gdp','mid']

Hyperparameter optimization

- To choose a set of optimal hyperparameters for a learning algorithm
- To control the learning process

Hyperparameter Tuning Methods

- Grid search

Grid search

- Model Parameters are placed in the form of a matrix
- Each set of parameters is taken into consideration and the accuracy is noted
- The model with the set of parameters which give the top accuracy is considered to be the best

Hyperparameter optimization

Example of Grid Search:

For Support Vector Machine parameters:

```
C = [1.0, 2.0, 5.0, 10.0]
kernal= ['rbf', 'linear', 'poly']
```

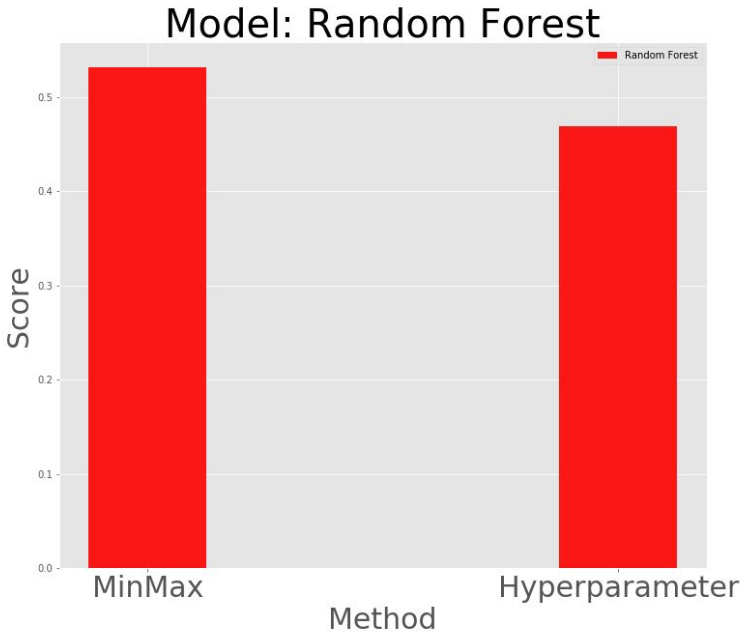
All combinations to be tested:

```
svm.SVC(C=1.0 kernel='rbf')
svm.SVC(C=1.0 kernel='linear')
svm.SVC(C=1.0 kernel='poly')

svm.SVC(C=2.0 kernel='rbf')
svm.SVC(C=2.0 kernel='linear')
svm.SVC(C=2.0 kernel='poly')

svm.SVC(C=5.0 kernel='rbf')
svm.SVC(C=5.0 kernel='linear')
svm.SVC(C=5.0 kernel='poly')
```

Hyperparameter optimization

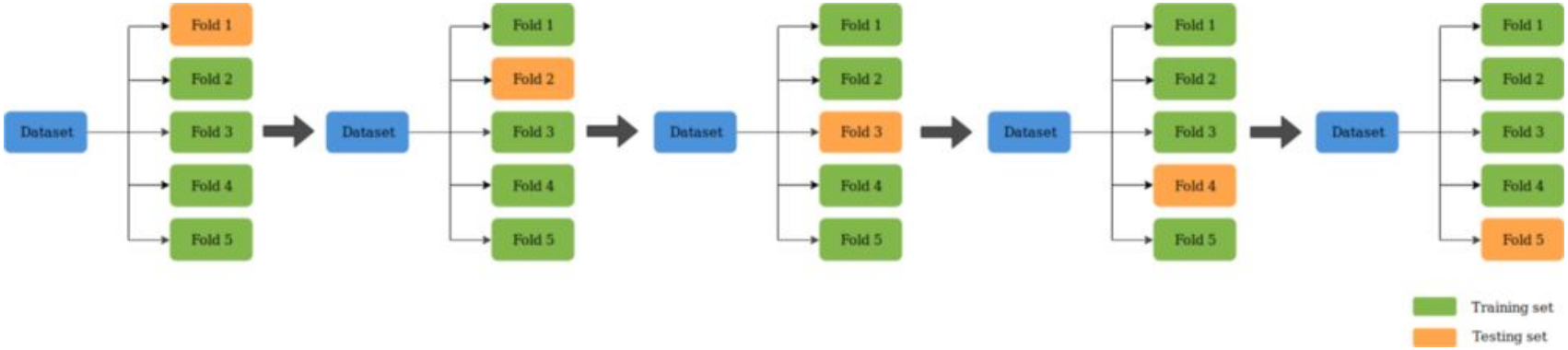


Cross Validation

- Evaluating model with fixed training and testing set is not reliable
- The accuracy obtained for one test set can be very different to the accuracy obtained for a different test set

Solutions

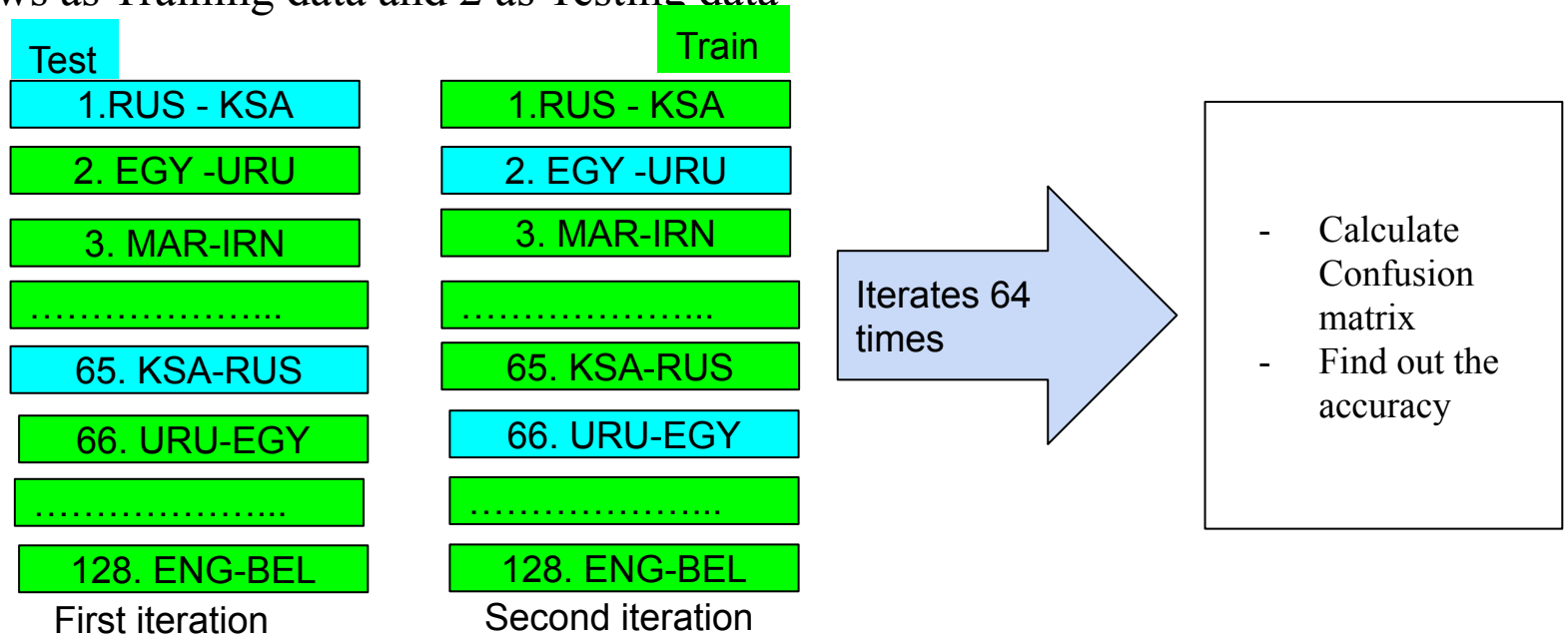
K-fold Cross Validation(CV)



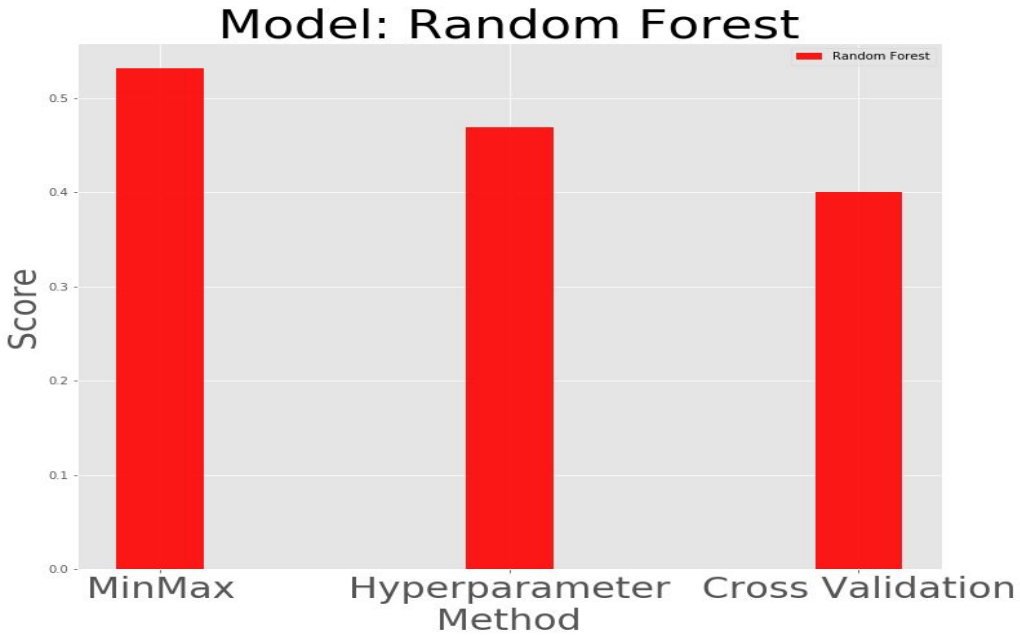
5-Fold Cross Validation
Image: [15]

Cross Validation

- 64 matches in total (With duplicate: 128)
- 126 rows as Training data and 2 as Testing data



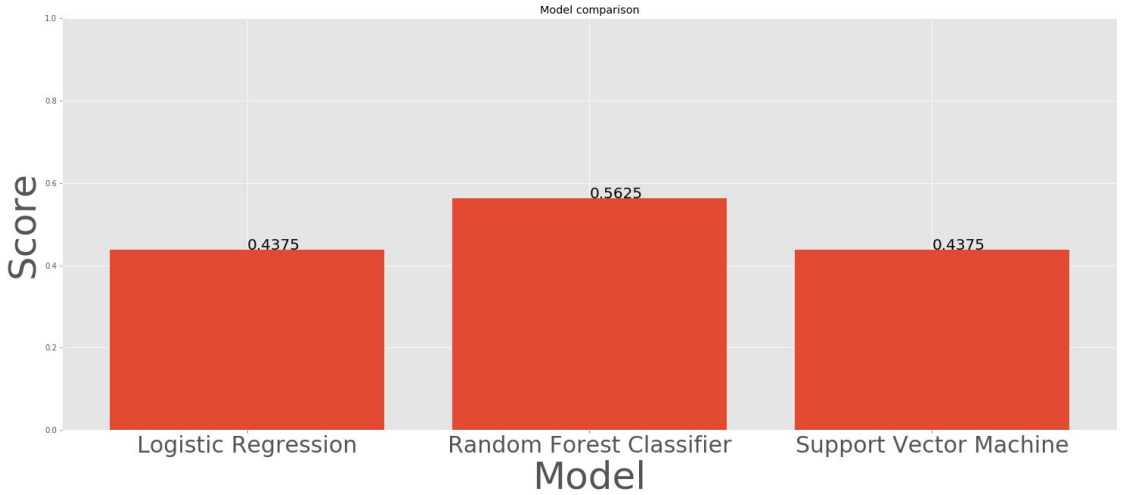
Cross Validation



Model Evaluation

Accuracy Score:

$$\texttt{accuracy}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} 1(\hat{y}_i = y_i)$$



Model Evaluation

Confusion Matrix: Binary Class

- Handling imbalanced class problem

		Predicted Results	
		Positive	Negative
True Condition	Positive	TP(True Positive)	FN(False Negative)
	Negative	FP(False Positive)	TN(True Negative)

Source: [18]

Model Evaluation

Confusion matrix: multiclass

true class \ inferred class	A	B	C
A	a	b	c
B	d	e	f
C	g	h	i

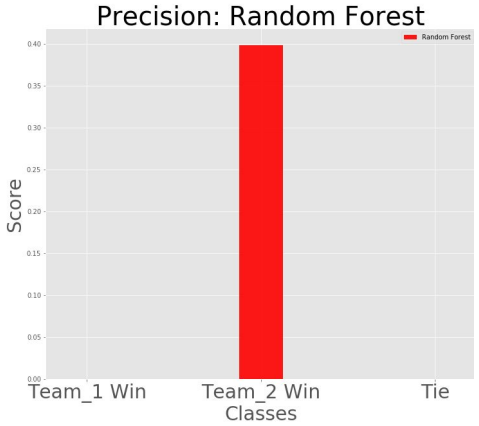
true class \ inferred class	A	not−A
A	a (TP)	b+c (FN)
not−A	d+g (FP)	e+f+h+i (TN)

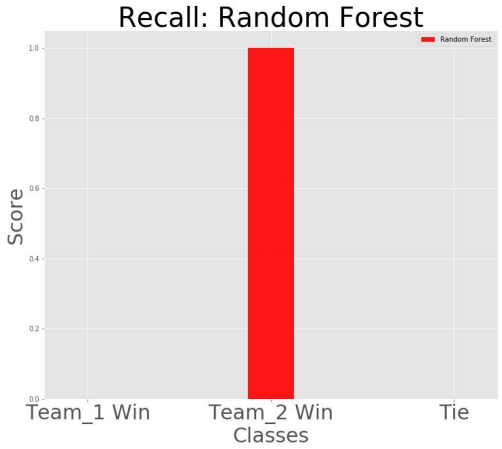
Model Performance

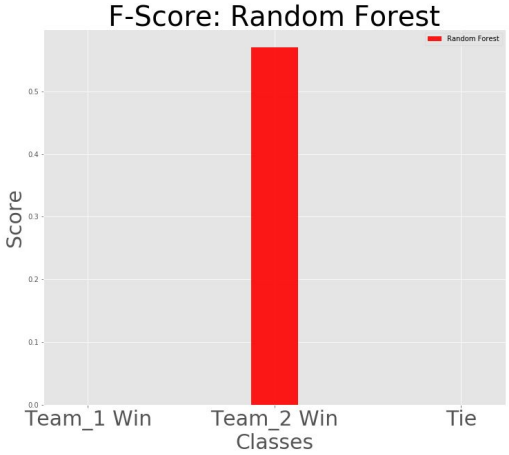
Accuracy	= (TN+TP)/(TN+FP+FN+TP)
Precision	= TP/(FP+TP)
Sensitivity	= TP/(TP+FN)

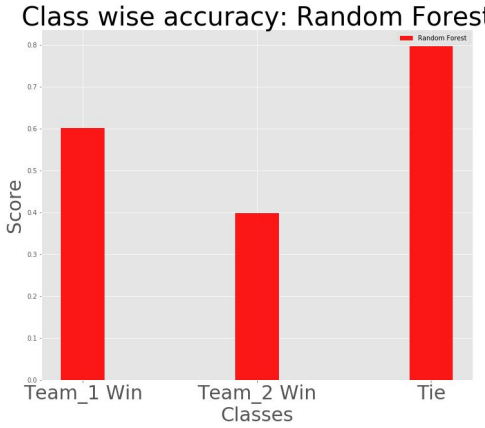
Source: [16] [17]

Model Evaluation

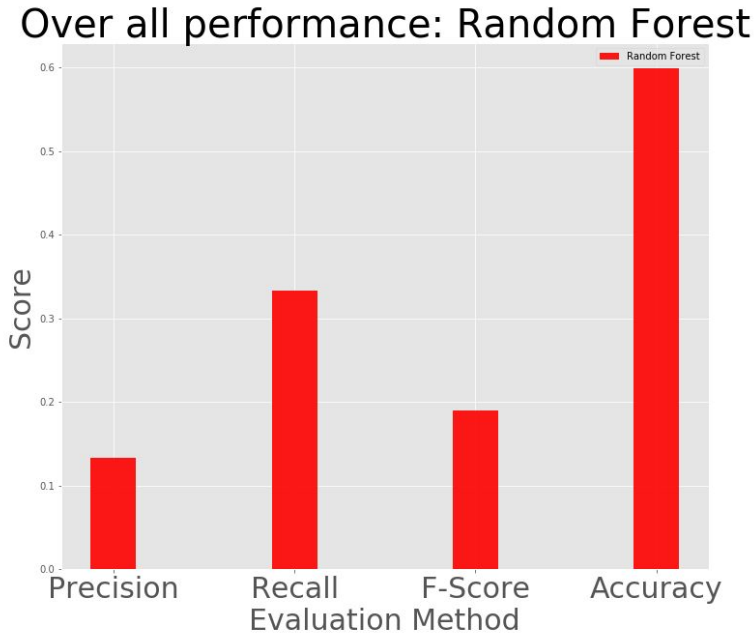








Model Evaluation



References

[1] https://en.wikipedia.org/wiki/List_of_countries_by_GDP_(PPP)_per_capita

[2] https://www.cia.gov/library/publications/resources/the-world-factbook/rankorder/2119rank.html

[3] https://data.worldbank.org/indicator/SP.POP.TOTL?view=chart

[4] https://worldhappiness.report/ed/2017/

[5] cover image: https://www.soundersfc.com/post/2018/05/15/svensson-torres-and-lodeiro-take-next-steps-toward-2018-fifa-world-cup-russia

[6] coach rank: https://realsport101.com/news/sports/football/ranked-all-32-managers-at-the-world-cup/

[7] Star count: https://www.easports.com/fifa/fifa-18-player-ratings-top-100

[8] FIFAINDEX: *https://www.fifaindex.com/*

[9] *SPI, OFF, DEFF:* https://projects.fivethirtyeight.com/2018-world-cup-predictions/?ex_cid=endlink

References

[10]https://eu.usatoday.com/story/sports/soccer/worldcup/2018/05/16/world-cup-2018-cristiano-ronaldo-lionel-messi-head-top-50-list/610929002/

[11] Features result: https://fixturedownload.com/results/fifa-world-cup-2018

[12] Picture scale: https://medium.com/greyatom/why-how-and-when-to-scale-your-features-4b30ab09db5e

[13] Hypermeter: https://en.wikipedia.org/wiki/Hyperparameter_optimization

[14]Hypermeter:https://www.analyticsindiamag.com/why-is-random-search-better-than-grid-search-for-machine-learning/

[15] Cross validation: https://medium.com/datadriveninvestor/k-fold-cross-validation-6b8518070833

[16] Confusion Matrix: https://www.sciencedirect.com/science/article/abs/pii/S0376635717301146

[17] https://scaryscientist.blogspot.com/2016/03/confusion-matrix.html?view=classic

[18] https://www.mdpi.com/1424-8220/19/5/1137

[19]http://benalexkeen.com/feature-scaling-with-scikit-learn/

Technische Universität Berlin

Demo