

Applying Data Skills to a Life-Long Passion

AN *IMPOSSIBLE* SPORTS PREDICTION PERSONAL PROJECT

TALHA A. SIDDIQUI

Background

Personal

- Born and raised in Pakistan
- Watching and playing cricket my entire life

Education

- UBC's Master of Data Science
- BS Information Systems

Professional

- Data Scientist, aDolus Inc.
- Data and Analytics Consultant, KPMG US

Why Cricket Prediction?

- Passion
- Fun
- Personal
- Different
- Data Available
- Fits the Machine Learning Paradigm




Breaking Down the Problem

- Tournament
- Venue
- Teams
- Players

Result
6th match, ICC Cricket World Cup at Nottingham, Jun 3 2019

Pakistan 348/8
England 334/9 (50 ov)

Pakistan won by 14 runs

PLAYER OF THE MATCH
 **Mohammad Hafeez**
Pakistan

Summary Scorecard Report Commentary Videos Coverage Statistics Table

⏏ Pakistan Innings (50 overs maximum)

BATSMEN		R	B	M	4s	6s	SR
Imam-ul-Haq	✓ c Woakes b Ali	44	58	92	3	1	75.86
Fakhar Zaman	✓ st †Buttler b Ali	36	40	63	6	0	90.00
Babar Azam	✓ c Woakes b Ali	63	66	83	4	1	95.45
Mohammad Hafeez	✓ c Woakes b Wood	84	62	99	8	2	135.48
Sarfraz Ahmed (c) †	✓ c & b Woakes	55	44	68	5	0	125.00

Keep Breaking

- Players
- Attributes
- Bat
- Bowl

Mohammad Hafeez

Pakistan

Full name Mohammad Hafeez

Born October 17, 1980, Sargodha, Punjab

Current age 39 years 13 days

Major teams Pakistan, Faisalabad, Faisalabad Wolves, Guyana Amazon Warriors, Kolkata Knight Riders, Lahore Lions, Lahore Qalandars, Melbourne Stars, Montreal Tigers, Peshawar Zalmi, Sargodha, St Kitts and Nevis Patriots, Sui Northern Gas Pipelines Limited

Playing role Allrounder

Batting style Right-hand bat

Bowling style Right-arm offbreak



Batting and fielding averages

	Mat	Inns	NO	Runs	HS	Ave	BF	SR	100	50	4s	6s	Ct	St
Tests	55	105	8	3652	224	37.64	6520	56.01	10	12	455	28	45	0
ODIs	218	216	15	6614	140*	32.90	8633	76.61	11	38	664	110	85	0
T20Is	89	86	8	1908	86	24.46	1643	116.12	0	10	196	51	25	0
First-class	210	365	15	12169	224	34.76			26	56			183	0
List A	340	337	20	11402	140*	35.96			17	75			144	0
T20s	274	260	23	5753	102*	24.27	4760	120.86	2	31	605	169	90	0

Bowling averages

	Mat	Inns	Balls	Runs	Wkts	BBI	BBM	Ave	Econ	SR	4w	5w	10
Tests	55	77	4067	1808	53	4/16	4/48	34.11	2.66	76.7	2	0	0
ODIs	218	177	7733	5400	139	4/41	4/41	38.84	4.18	55.6	1	0	0
T20Is	89	67	1117	1226	54	4/10	4/10	22.70	6.58	20.6	1	0	0
First-class	210		14992	6764	253	8/57		26.73	2.70	59.2		7	2
List A	340		13269	9304	256	4/23	4/23	36.34	4.20	51.8	4	0	0
T20s	274	203	3773	3994	174	4/10	4/10	22.95	6.35	21.6	4	0	0

Data Science Workflow

Data Scrapping

- 6,000+ web pages: 4,000 one-day international matches played over 50 years by over 2,000 cricketers

Python Programming

- Reproducible scripts and Jupyter Notebooks

Data Visualization

- Python matplotlib and R ggplot2

Machine Learning

- Scikit-learn models

Project Management

- Issues / Projects

Documentation

- Wiki

The screenshot shows a GitHub repository page for 'talhaadnan100 / 2019-ICC-Cricket-World-Cup-AI-Predictions'. The repository has 22 commits, 1 branch, 0 releases, and 1 contributor. The main branch is 'master'. The repository description is 'Using AI to predict the outcome of the 2019 ICC Cricket World Cup'. The repository contains several files and folders: 'data' (Compiled Data Files), 'img' (Initial prediction notebook created), 'notebooks' (Initial prediction notebook created), 'scripts' (Updated player details aggregator and running script from terminal), 'LICENSE' (Initial commit), and 'README.md' (Initial prediction notebook created). The 'README.md' file is open, showing the title '2019 ICC Cricket World Cup AI Predictions' and the description 'Using AI to predict the outcome of the 2019 ICC Cricket World Cup'.

talhaadnan100 / 2019-ICC-Cricket-World-Cup-AI-Predictions

Watch 0

Code Issues 1 Pull requests 0 Projects 0 Wiki Security Insights Settings

Using AI to predict the outcome of the 2019 ICC Cricket World Cup

sports-analytics predictive-modeling sklearn webscraping python3 Manage topics

22 commits 1 branch 0 releases 1 contributor

Branch: master New pull request Create new file Upload files

talhaadnan100 Initial prediction notebook created

data	Compiled Data Files
img	Initial prediction notebook created
notebooks	Initial prediction notebook created
scripts	Updated player details aggregator and running script from terminal
LICENSE	Initial commit
README.md	Initial prediction notebook created

README.md

Work in Progress

2019 ICC Cricket World Cup AI Predictions

Using AI to predict the outcome of the 2019 ICC Cricket World Cup

Predictions

```
data = pd.read_csv("../data/matches_scorecard_player_details.csv")
print("Shape:", data.shape)
#print(list(data.columns))
```

Shape: (4010, 277)

```
# Random Forest
param_grid = {'n_estimators' : [10, 50, 100],
              'criterion' : ['gini', 'entropy'],
              'max_depth' : [10, 50, 100],
              'min_samples_split': [2, 5, 20]}

rfc_gridsearch = GridSearchCV(estimator=RandomForestClassifier(), param_grid=param_grid)
print("Number of configurations:", np.prod(list(map(len, param_grid.values()))))

rfc_gridsearch.fit(X_train, y_train)
print('Best Parameters', rfc_gridsearch.best_params_)
print("Training Accuracy:", rfc_gridsearch.score(X_train, y_train))
print("Test Accuracy      :", rfc_gridsearch.score(X_test, y_test))
```

Number of configurations: 54

Best Parameters {'criterion': 'gini', 'max_depth': 10, 'min_samples_split': 20, 'n_estimators': 100}

Training Accuracy: 0.9281676089125375

Test Accuracy : 0.6630109670987039

Predictions

```
# Feed forward Neural Network (MLP Classifier)
param_grid = {'hidden_layer_sizes' : [(50,), (10,), (10,10)],
              'learning_rate_init' : [1e-4, 1e-3, 1e-2],
              'alpha' : [1e-5, 1e-4, 1e-3],
              'activation' : ['relu', 'tanh']}

mlp_gridsearch = GridSearchCV(estimator=MLPClassifier(), param_grid=param_grid)
print("Number of configurations:", np.prod(list(map(len, param_grid.values()))))

mlp_gridsearch.fit(X_train, y_train)
print('Best Parameters', mlp_gridsearch.best_params_)
print("Training Accuracy:", mlp_gridsearch.score(X_train, y_train))
print("Test Accuracy      :", mlp_gridsearch.score(X_test, y_test))
```

Number of configurations: 54

Best Parameters {'activation': 'tanh', 'alpha': 1e-05, 'hidden_layer_sizes': (50,), 'learning_rate_init': 0.0001}

Training Accuracy: 0.7273029597605587

Test Accuracy : 0.6520438683948155

Predictions

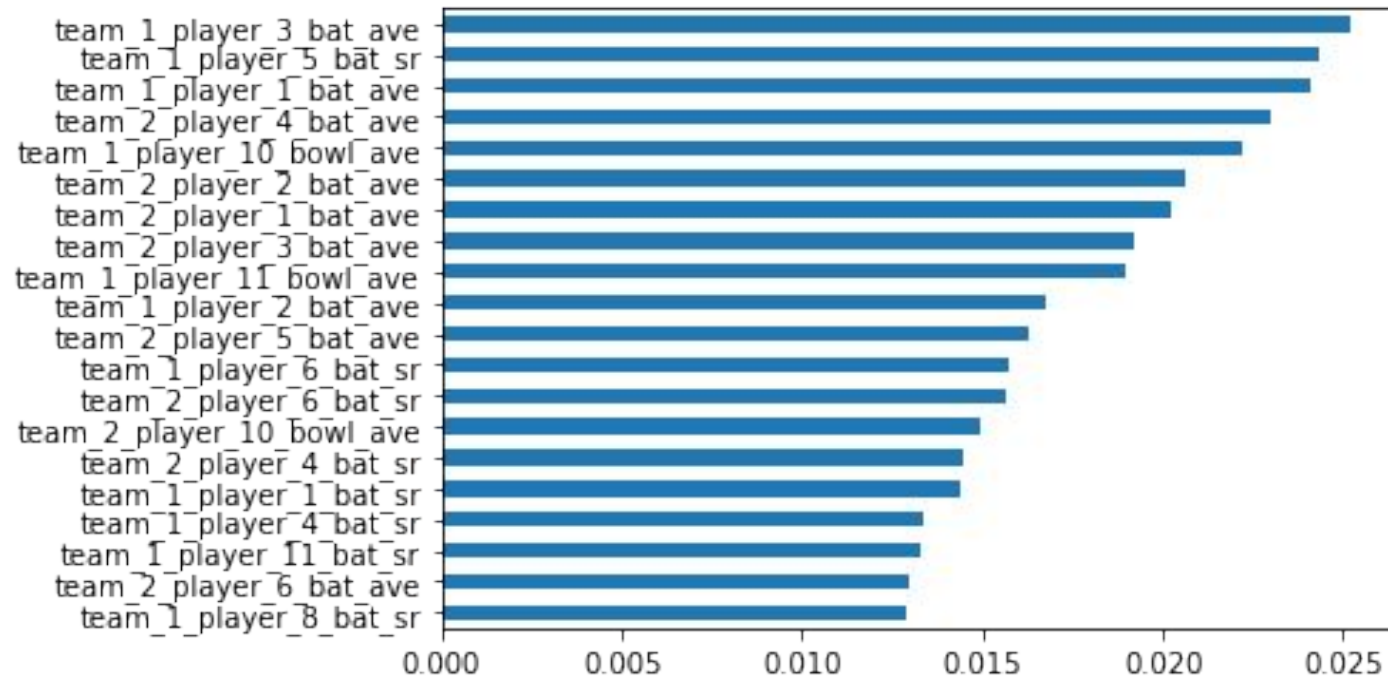
```
# SVC
param_grid = {'C' : [0.001, 0.01, 0.1, 1, 10],
              'kernel' : ['rbf', 'poly', 'sigmoid', 'linear'],
              'gamma' : [0.001, 0.01, 0.1, 1]}

svc_gridsearch = GridSearchCV(estimator=SVC(), param_grid=param_grid)
print("Number of configurations:", np.prod(list(map(len, param_grid.values()))))

svc_gridsearch.fit(X_train, y_train)
print('Best Parameters', svc_gridsearch.best_params_)
print("Training Accuracy:", svc_gridsearch.score(X_train, y_train))
print("Test Accuracy      :", svc_gridsearch.score(X_test, y_test))
```

```
Number of configurations: 80
Best Parameters {'C': 1, 'gamma': 0.001, 'kernel': 'sigmoid'}
Training Accuracy: 0.6923844363152644
Test Accuracy      : 0.6630109670987039
```

Feature Importance



Lessons Learnt

- Did I succeed?
- Skills to showcase
- Quality over quantity
- Passion
- More about the data, less about the model
- Make it public
- Start small

How to Get Started?

- [R for Data Science's #TidyTuesday](#)
 - David Robinson's Tidy Tuesday Screencast
- [FiveThirtyEight](#)
- [Kaggle Datasets](#)
- [UCI Machine Learning Repository](#)
- [Namara.io](#) by ThinkData Works
- Open Data Portals by [City of Vancouver](#), [BC Government](#), [StatisticsCanada](#)

Questions?

bit.ly/cricket-data-project