

Master of Technology (IS)

FIFA 19 Classification Project

Project Report

Team Member:

CAO LIANG A0012884E

GENG LIANGYU A0195278M

Contents

[Executive Summary 3](#_Toc20067869)

[Business Problem Background 3](#_Toc20067870)

[Project Objectives & Success Measurement 3](#_Toc20067871)

[Project Solution Design 4](#_Toc20067872)

[Project Files Structures 5](#_Toc20067873)

[Project Implementation 6](#_Toc20067874)

[1. Row Data Collection 6](#_Toc20067875)

[2. Pre-Processing 6](#_Toc20067876)

[3. Sampling 8](#_Toc20067877)

[4. Training Learning Algorithms 11](#_Toc20067878)

[4.1 Logistic Regression 11](#_Toc20067879)

[4.2 Decision Tree 12](#_Toc20067880)

[4.3 Neural Network 13](#_Toc20067881)

[4.4 Random Forest 14](#_Toc20067882)

[4.5 Hybrid Classifier 15](#_Toc20067883)

[4.6 Additional Approach 16](#_Toc20067884)

[5. Hyper-parameter Optimization 17](#_Toc20067885)

[6. Post-Processing 17](#_Toc20067886)

[7. Final Classification Model 21](#_Toc20067887)

[Project Performance & Validation 22](#_Toc20067888)

[Project Conclusions: Findings & Recommendation 23](#_Toc20067889)

# Executive Summary

FIFA, also known as FIFA Football or FIFA Soccer, is a series of association football video games or football simulator, released annually by Electronic Arts under the EA Sports label. FIFA 19 data set comes from “<https://www.kaggle.com/karangadiya/fifa19>” which gets player information from “<https://sofifa.com/>”.

This project utilizes the several sklearn classifiers algorithms to predict the FIFA 2019 player international reputation scores.

# Business Problem Background

EA Sports’ expensive and luxurious FIFA 19 is undoubtedly football games’ equivalent of the Premier League. The international reputation values the football game player, and it is important factor to consider when choosing game player. Therefore, the project provides ways to predict and realise the importance of international reputation score.

# Project Objectives & Success Measurement

We will utilize existing FIFA 19 dataset and predict international reputation scores with selected relevant feature set.

The measurement for the project will depends on the prediction accuracy on the test data, and it should achieve 80% average accuracy or more to be success.

# Project Solution Design

The project should prepare the dataset with sample size greater than 3000, then select the sklearn classifiers and conduct the training, finally perform the verification of the training result according to average accuracy.

The project follows the supervised learning process as below.

# Project Files Structures

Current project contains following files:

1. conda-install.txt, packages which are needed by this project.
2. classification-overall.ipynb, file which contains all the algorithms being used.
3. classification-eli5\_hyperopt.ipynb, an additional approach based on eli5 and hyperopt.
4. prepare-data.ipynb, the code we used to clean up and prepare the ready-to-use data.
5. features-elimination.ipynb, code used to eliminate unnecessary features.
6. /py\_src, contains all the python source files, which can be run from PyCharm or Spyder, common\_functions.py contains the functions which are shared by all the other files.

The other files are named by according algorithms used inside.

1. fifa19\_ready\_data.csv, processed data which is being used by all the algorithms.
2. raw\_data.csv, raw dataset gets from Kaggle.

# Project Implementation

The project performs the steps below.

## Row Data Collection

The Row Data “data.csv” contains 18208 records, and we select the records randomly to get 4000 records, so that there are extra records which can be filtered during pre-processing stage, and we still can meet minimum 3000 records requirement for the project.

## Pre-Processing

There are 4000 records with 89 features in the beginning of this stage, i.e. shape (4000, 89). We have spent lots of efforts to perform the following to process the records, so they are ready for training.

* 1. Removing Invalid Data

We perform this step so the overall data accuracy can be higher.

* + 1. Since we have set the target on “International Reputation”, so the records with missing “International Reputation” value will be removed.
    2. The scores for the root player positions (i.e. 'RAM', 'RB', 'RCB', 'RCM', 'RDM', 'RF', 'RM', 'RS', 'RW', 'RWB', 'ST') must exist. So if these score values are missing, the records also are removed.

The root positions are chosen with the following reason.

Since the player position left, centre and right are all considered the same score, so we pick only right player position to represent all positions (left, centre and right), and the right player position is called root position.

After 2-1 step, the records shape is reduced to (3563, 89).

* 1. Removing Features

We perform the steps below to increase the accuracy and also reduce the processing since the data shape will be reduced a lot.

* + 1. ‘Row\_number’ is removed since it only reflects the original data row number, and will not have any meaning for the project.
    2. The following features are removed since they are not used in the project.
* 'Name', 'Photo', 'Flag', 'Club Logo', 'Real Face', 'Jersey Number', 'Joined', 'Loaned From', 'Contract Valid Until', 'Release Clause'
  + 1. The following features are removed since we consider to choose the features without any bias. E.g. the international reputation score should not be related to Nationality, so people from different countries will be treated fairly regarding to international reputation.
* 'Nationality', 'Club', 'Preferred Foot', 'Body Type', 'Position', 'Weak Foot'
  + 1. The following player positions are removed since they have been represented by the root player positions.
* 'CAM', 'CB', 'CDM', 'CF', 'CM', 'LAM', 'LB', 'LCB', 'LCM', 'LDM', 'LF', 'LM', 'LS', 'LW', 'LWB'

After 2-2 step, the records shape is reduced to (3563, 57).

* 1. Converting Values

The values inside the data set contain many invalid data, such as empty value and string value with different format. We perform the steps below to convert the string value to numeric values, and also fill in the adequate numeric value for missing value.

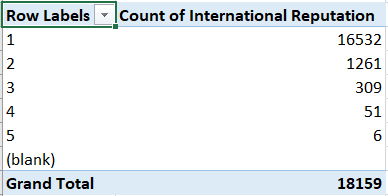
* + 1. Convert market and wage to the numeric value, and fill in mean of market and wage values individually in case of missing value.
    2. Convert height value from mixed inch and foot string value to the numeric value in foot only. Also fill in missing height value with mean height value.
    3. Convert weight value from string to numerical value, and fill in missing value with mean weight value.
    4. Convert position string value to numerical value.
    5. Convert work rate value to separate attack value and defence value in number. Work rate value is string value which contains attack rate and defence rate category, e.g. Medium/ Medium, and it is converted to numerical value, i.e. High is 3, Medium 2, Low 1.
    6. Convert the following string values to numeric values, and fill in mean value on the missing value.
* 'Age', 'Overall', 'Potential', 'Special'

After pre-processing, the data set shape is (3563, 32), and is reduced a lot compared to original shape (4000, 89). Also it meets the project requirement for at least 3000 records with 30 features.

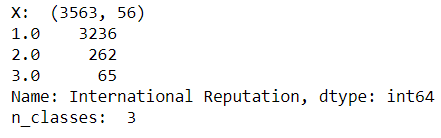
## Sampling

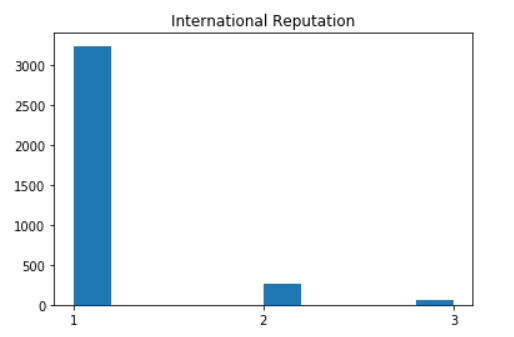
* 1. Adjusting Rating Score

Since rating score 5 has only 6 records within all 18159 records, and rating score 4 has only 51 records, we choose the re-assign the records with rating score 5 or 4 to be rating score 3, so the accuracy will be reasonable.



After that, the international reputation scores distribution is shown below.

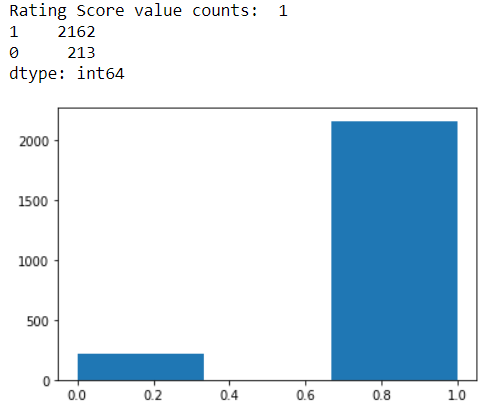


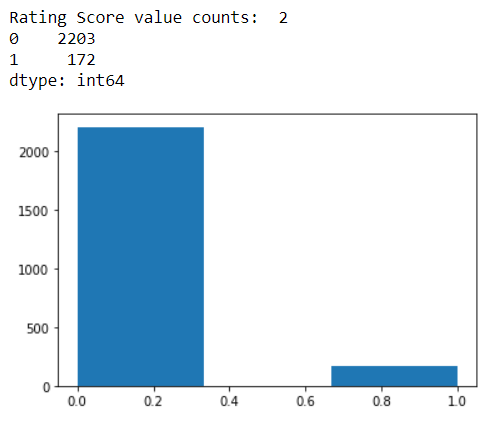


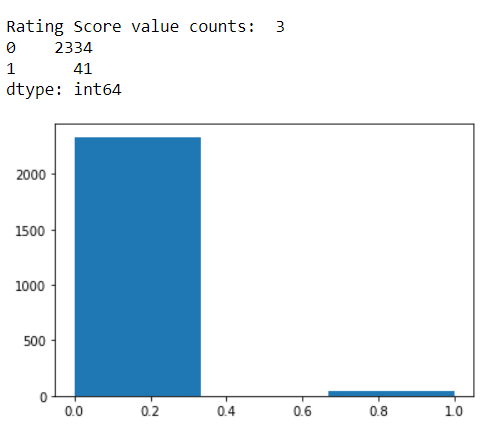
* 1. Converting Rating Score Label to Binary Categories Values

We utilize ‘label\_binarize’ function to convert rating score values to 3 binary categories values. The data distribution for different rating scores are shown below.

This step will enable the one vs rest classifiers at training stage can be utilized. And from our experiment, this improves the accuracy result significantly, when comparing the way if we create 3 different classifiers in same type and then train them.







* 1. Splitting Training Data Set and Testing Data Set

We split data set as 2/3 for training, and the rest 1/3 for testing.

* 1. Scaling Data

We scale data with ‘StandardScaler’ before training.

## Training Learning Algorithms

We have implemented 6 classifiers to deal with the classification problem.

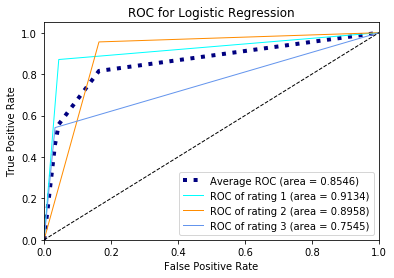
Since the evaluation measure for multi-class classification is macro-averaging, which gives equal weight to the classification of each label. So, the final average accuracy might be a bit different from the ROC chart displayed.

## Logistic Regression

We choose the option multi\_class='ovr', and also use OneVsRestClassifier to improve the overall accuracy.

After optimization with GridSearchCV, the accuracy can reach 0.9153.

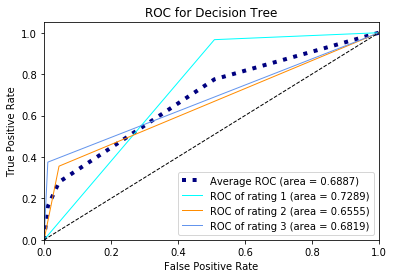
ROC chart as below:



## Decision Tree

We use OneVsRestClassifier to improve the accuracy. The final accuracy score reaches 0.8726 after GridSearchCV optimization.

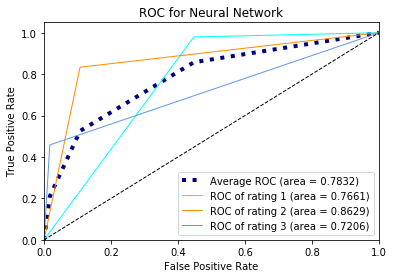
ROC chart as below:



## Neural Network

We use OneVsRestClassifier to improve the accuracy, and the final accuracy score is 0.8903 after GridSearchCV optimization.

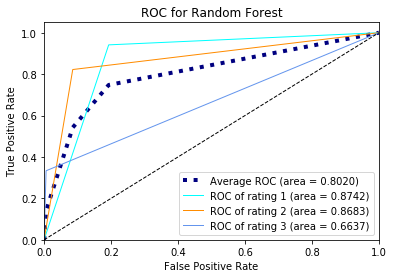
ROC chart as below:



## Random Forest

We also add OneVsRestClassifier and the final average accuracy score is 0.9131 after GridSearchCV optimization.

ROC chart as below:

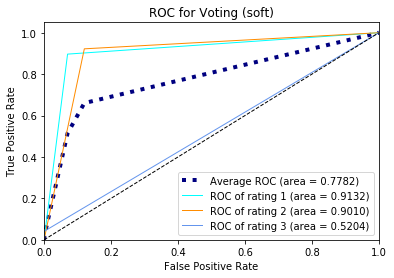


## Hybrid Classifier

Due to OneVsRestClassifier is not supported by sklearn.ensemble.VotingClassifier, so we have implemented a customized HybridClassifier. So, it performs the voting with multiple classifier, and supports OneVsRestClassifier multiclass classification.

We applied it on all top 2 classifiers, and the final average accuracy is 0.9166.

ROC chart as below:



## Hyper-parameter Optimization

GridSearchCV is being used regarding parameter tuning.

We are able to find better parameters with the functions provided by it.

Besides above GridSearchCV, we also tried an additional approach to optimize MLP hyper parameters using following techniques:

eli5: Use for feature importance determination through permutation.

hyperopt: used for optimizing over awkward search spaces with real-valued, discrete, and conditional dimensions.

This model keeps updating its beliefs on the best hyperparameter combination, which was implemented in hyperopt.

The state space for each of the seven parameter:

Activation: logistic, tanh, relu, identity

Alpha: float between 0.001 to 1000

Neurons: integer between 1 to 55 (i.e., number of features)

Learning rate: constant, invscaling, adaptive

Initial: float between 0.001 to 0.99

Exponent: float between 0.01 to 0.99

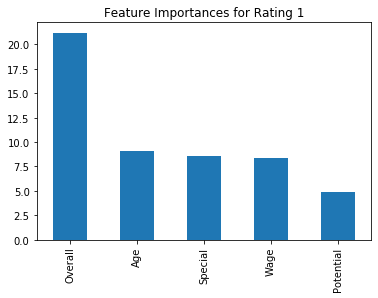
Solver: lbfgs, sgd, adam

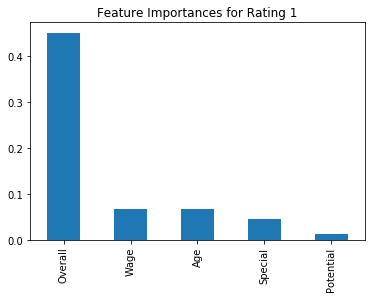
Details can be found in the file: classification-eli5\_hyperopt.ipynb.

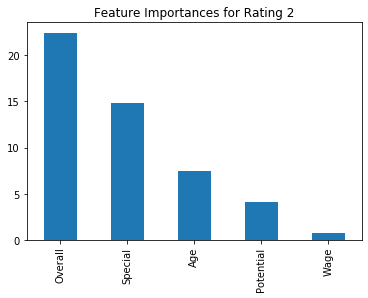
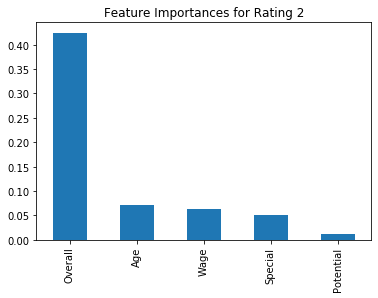
## Post-Processing

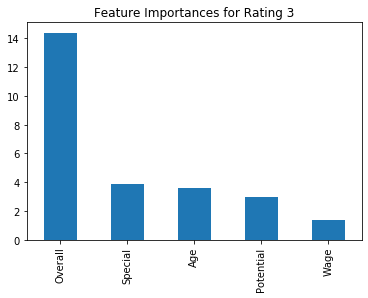
After building on the voting hybrid classifier, we learn the feature importance based on the 2 estimators used inside: Logistic Regression Classifier, and Random Forest Classifier, so to better understand the business value of the classification.

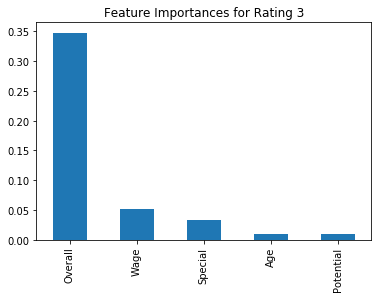
The top 5 most important features from 2 different estimators are same, and they are listed below. And it shows the most important feature is Overall, and it outweighs other features.









## Final Classification Model

We select the best performance model which is Voting hybrid classifier as the final classification model.

Followings are the accuracy result summary for each model.

Voting Hybrid Classifier, score: 0.9166

Logistic Regression, score: 0.9153

Random Forest, score: 0.9131

Neural Network, score: 0.8903

Decision Tree, score: 0.8726

# Project Performance & Validation

Base on the accuracy from different models, the best testing average accuracy we can get is 91.66% from voting hybrid classifier, and this achieves the project performance requirement that requires the average accuracy to be greater than 80%. The average accuracy 91.66% is tested with the testing data set, so this validates the result.

# Project Conclusions: Findings & Recommendation

For the players with international reputation which are 2 and 3, the results are not very good due to insufficient data. We practiced with other scikit learn classes and some other ways to approach the accuracy, not just the classes used in workshops.

In the future, we will try to grab more data from FIFA regarding the players which international reputation rating are 2 and 3, that would help increase our prediction accuracy.

We also learn that the pre-processing is important to clean up invalid data, and all the classifier can be optimized with hyper-parameter tuning, and finally the ensemble/hybrid classifier can improve the performance with 2 or more classifiers.