Sadhana Singh

Classification model used is: K Nearest Neighbors

Brief Steps:

Step1: Read the csv file

Step2: Feature Selection: Select appropriate features

Step3: Data munging: Selecting subset of data that is useful is predicting rather than complete data

Step4: Partition the dataset: Partition into 75% training and 25%test data

Step5: Train the dataset: Apply the KNN with best parameters

Step6: Predict: Predict classes for test dataset

Step7: Accuracy: Calculate the accuracy

Step8: Confusion Matrix: Print the confusion Matrix

Step9: Cross Validation: Apply KNN with same parameters, use 10 fold validation

Step10: Average Cross Validation: Calculate Average of all Cross validation values

**Details:**

1. I have implemented KNN Model for classification. Initial: With all features on given dataset and simple implementation of KNN with default parameters gave me accuracy of 36%

**Further I have done feature selection, Data cleaning and Data preprocessing that improved the test accuracy to 72%**

1. I selected 8 features that were derived from feature\_importance module of decision tree model. I have provided the code for it in the .py file. However, I have commented it and hard coded it in feature\_columns\_selected as for every run, it generates list dynamically that has variance in accuracy.

I have selected top 8 features based on the weight of the features and also on some Domain Knowledge.

Below is some domain knowledge that I applied in feature selection:

I have used these feature =['3PA', 'AST','TRB', 'eFG%', 'BLK', 'STL', 'FT%', 'ORB']

3PA being 0, is highly indicative of Centre Position

PG are team’s best handlers and passer. ASSISTS and STEALS is highly indicative for thier Postion. Also they are quick

SG are good shooters and play 3P range

Rebounds if more gives more information about Centre. Hence included Offensive and Total Rebounds. Centre players are typically skilled at gathering rebounds

eFG% - It is one of most important factor in consideration as it states the fact that 3P goal is worth one more point than 2P goal.

FT% - Free throws are unopposed attempts from restricted area on the court. It also defines the position.

BLK – Centres and [Power](https://en.wikipedia.org/wiki/Power_forward_%28basketball%29) Forwards tend to record the most blocks, hence help in classifying their position

The feature importance functionality provided similar results and helped in the feature selection

1. I have performed Data cleaning based on some domain knowledge.

The Players that have played less number of minutes do not provide much information. Also the data with player that play less number of games are not much useful

So, I have selected data of players with minutes greater than 4 and games greater than 4. So it trains on subset of data that has players with more experience that that have played more than 4 minutes in more than 4 games.

Block is one of the important feature and it is mainly done by Centres and PF, remove the outliers and omitted players that played block more than 2.3

1. I have used Exhaustive Grid Search for parameter selection in KNN model. I have commented this code in the .py file due to the dynamic change in the parameters that leads to variance in accuracy.

I have implemented the best parameter based on multiple runs and best combination and accuracy results.

1. Normalize: I have further normalized the values to standardized the data. It scales individual samples to have unit norm
2. I have trained the cleaned or processed normalized data and tested on complete data set. I have normalized this test data.
3. Further printed the Test accuracy and confusion Matrix
4. For Validation, since cross\_val\_score does not provide separate provision to train on one subset of data and test on complete set of data.

I have implemented cross validation on my own as well.

I have used two types of data set here, one for training and later for testing purpose

The Data set for training purpose is subset of complete NBAStats and includes only those players that have played for more than 4 minutes and played more than 4 games

The Data set for testing purpose is complete data.

For both datasets, I am normalizing the values to standardize it.

References:

<https://en.wikipedia.org/wiki/Basketball_statistics>

<http://www.basketball-reference.com/leagues/NBA_2016_per_game.html>

<http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html#sklearn.neighbors.KNeighborsClassifier>