**Algorithm**

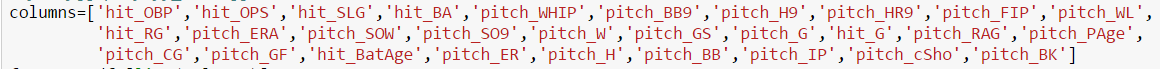
1). Logistic Regression

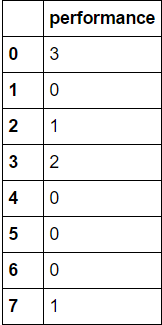
2). KNN Model

Logistic Regression

We imported logistic regression from scikit learn package.

Input Variables- We used the top 29 features as input variables using Recursive Feature Elimination



Output - We divided the team performance into four categories [0,1,2,3,4] where

0->implies team lost in Division Series

1-> team lost in Champions Series

2-> team is World Series loser

3->Team won the World series

Step1

The data was split into training and test variables as

x\_train->It contains input variables data from year 1995-2015 for all playoff teams

y\_train->It contains output data which has performance of all teams from year 1995-2015

x\_test->Input variable data for year 2016

y\_test-> performance of all teams in year 2016

Step 2

Fit the training input(x\_train) and output data(y\_train) into logistic regression model to train the algorithm

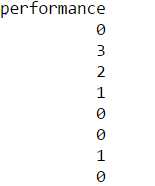
Step 3

Predict the output performance(y\_pred) of all eight teams for year 2016 with input(x\_test)

Step 4

Compare the predicted output performance(y\_pred) with actual performance(y\_test) to get various parameters like accuracy, precision, recall and F1 score.

Conclusion- We were able to predict the world series winner correctly using this model.



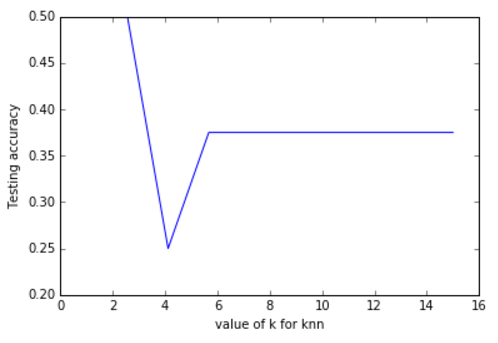
Actual Performance Predicted Performance

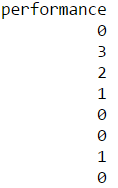
Accuracy-62%, Precision-52.5%, Recall-56.25% and F1 Score=54.166%

KNN Model

We applied the K Nearest Neighbor classifier on above data while splitting into same x\_train, y\_train, x\_test and y\_test variables

On plotting the graph between k-nearest neighbor and Testing Accuracy we found that accuracy is higher when k=1 or 2



The output result of KNN model is



Predicted Performance

Actual Performance

Accuracy-50%, Precision-50%, Recall-50% and F1 Score=50%

Conclusion- We are able to predict some of the teams which were world series loser, Division series loser, Champion series loser but the model has a shortcoming of not able to predict the world series winner for year 2016.

Some other variations we tried on dataset

1). We applied the KNN and Logistic Regression by dividing the training and test data set into 60:40 ratio to predict the performance of team for multiple years

KNN Output

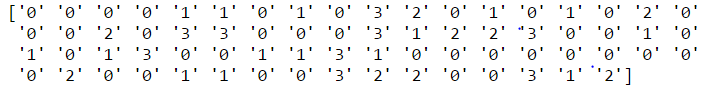
Accuracy 0.342857142857

Precision 0.294429050356

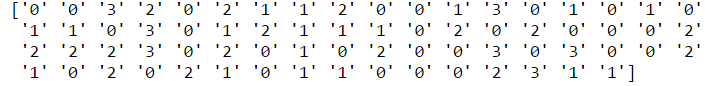
Recall 0.274774774775

F1 score 0.27946969697

Actual Performance



Predicted Performance



Logistic Regression Output

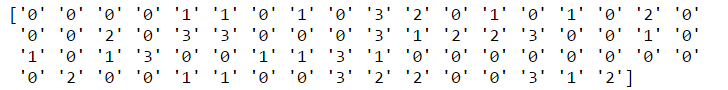
Accuracy 0.342857142857

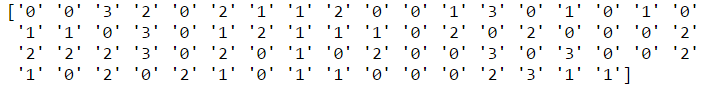
Precision 0.294429050356

Recall 0.274774774775

F1 score 0.27946969697

Actual Performance



Predicted Performance

2). KNN and Logistic Regression with K-mean cross validation by splitting Training and Test data into 60:40

We applied the KNN and logistic Regression using k-mean cross validation to check the accuracy and precision

KNN output-> Accuracy=50.95% Precision=27.87%

Logistic Regression -> Accuracy=49.20% Precision=24.20%