|  |  |
| --- | --- |
| **Assignment details** | |
| **Course** | ISCG 8043: ADAPTIVE BUSINESS INTELLIGENCE |
| **Assignment name** | Assignment 2: |
| **Due date** | 7/10/2016 |
| **Semester** | 2 |
| **Lecturer** | SIRA YONGCHAREON |

|  |  |
| --- | --- |
| **Student Name** | **Student ID** |
| DEEPANKAR SAJWAN | 1454355 |

Contents

[1. INTRODUCTION AND PROBLEM IDENTIFIED 3](#_Toc455881051)

[1.1 INTRODUCTION 3](#_Toc455881052)

[1.2 PROBLEM IDENTIFIED 4](#_Toc455881053)

[2. TECHNOLOGY 4](#_Toc455881054)

[2.1 CLASSIFICATION 4](#_Toc455881055)

[2.2 DECISION TREE 6](#_Toc455881056)

[2.3 RANDOM FOREST 7](#_Toc455881057)

[2.4 ADVANTAGES 7](#_Toc455881058)

[2.5 LIMITATIONS 8](#_Toc455881059)

[3. SOLUTION DESIGN AND DEVELOPMENT 9](#_Toc455881060)

[3.1 DATA PREPARATION MODULE 10](#_Toc455881061)

[3.2 DATA CLEANING/CORRECTING 11](#_Toc455881062)

[3.3 DATA TRANSFORMATION 11](#_Toc455881063)

[3.4 TRAINING OF MODEL 13](#_Toc455881064)

[4. EVALUATION 17](#_Toc455881065)

[5. FUTURE DEVELOPMENT AND CONCLUSION 18](#_Toc455881066)

[REFERENCES 19](#_Toc455881067)

# INTRODUCTION AND PROBLEM IDENTIFIED

* 1. ***Introduction***

With the modernization in the technology and information period managers need to distinguish the various competitive chances characterized by most of the decision support tools. Combination of prediction and also various optimization methods to help in quickly changing and complex surroundings. Various essential component or basics from predictive modelling, optimization, adaptability, forecasting and data mining has been included ABI (Adaptive Business Intelligence).

Adding further to this, huge amount of data has been stored by numerous business managers in a trust that they are holding something valued insight. However, the data stored by managers is raw and there is hardly any profit from it, and to get outcome from them it totally depends on the organization capability of analyzing this data. Then it gives birth to requirement of several software which can be skillful of interpreting, retrieving, summarizing data for users. And further to extract data and knowledge from raw many companies linked with business intelligence come in existence which specialized and dedicated for providing services and system related software’s. Information has been provided in the form of tables, pies, charts, graphs, and statistics by different BI software by analyzing the functional data of organization. However most of the system in business intelligence has few general objectives like using different sources, accessing the data. Adding further to this information has been collected by data transforming and also it arranges simple and easy graphical interface to view this knowledge or information. In a common man language, we can also that business intelligence is something which includes gathering and processing of data and presenting this knowledge into a friendly method. Also which improves and enhances the decision making capability.

In this report we are mainly going to use one of the technique or technology of Adaptive business intelligence for making prediction of sport. Predicting the results for sport games has never been simple because it includes various issues and which also got affected by each other (Leung and Joseph, 2014). Various approaches have been experimented in past few years and ABI has been evolved with great demand. Author in the paper discusses about classification technique used in basketball prediction (Haghighat, Rastegari, & Nourafza, 2013). In old times or traditionally sports industry was not so wide spread and even predictions on the sports were given by various experts who only belong to sports itself.

***1.2 Problem Identified***

In this report the problem which we identified is mainly based on the prediction and standings of winning team in the league. We are going to make prediction on the basis of various algorithms and latest tools. Sport which we have chosen is ‘Basketball’. The problem here was to find the one of the best NBA team throughout the season. However, the prediction made in this report is based on the winner of NBA (National Basketball Association) league and also ranking of the teams in this league. NBA is a professional men’s basketball league mainly based in Northern America. In this reporting we are focusing on using ABI methodologies with information and match data of 2013-14 NBA season and standings of 2014 for predicting the outcomes of 2015 league (Cao, 2012).

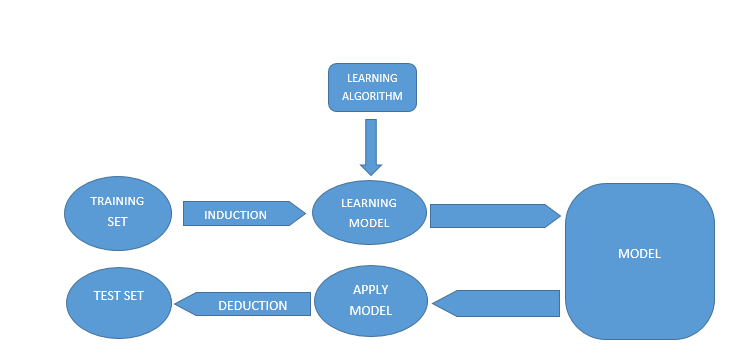
.

# TECHNOLOGY

The major aim of this report is to construct a prediction using the technology and method from ABI with highest accuracy possible. Adding further to this we are mainly focused on classification as major technology of ABI and further the method of classification which are Random forest and Decision tree has been functioning here for outcomes

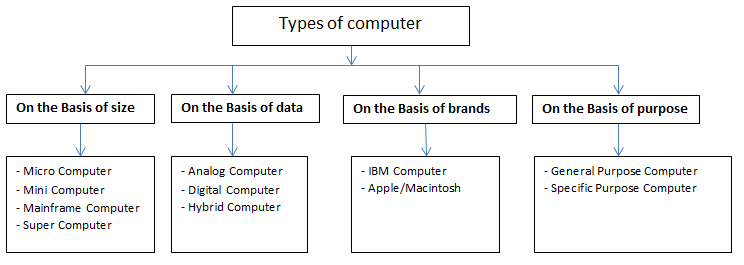
## ***Classification***

Classification is a process which can be defined as a technique of evolving functions with a strength to define the concept and data classes which usually depend on recognized characteristics and also within a separate dataset group it is been labelled. In supervised learning the method of characterization used is generally referred to classification. However, classification can also be defined with respect to data analyst and researchers who determine the class of unidentified trait which use a set of trained set. In ABI, Data classification act one of the important and significant role. Various application in classification are like computer diagnostics, medical diagnostics, troubleshooting, managing performance, prediction, forecasting. The figure below explains this.



*Figure 1*: Classification Model

The outcome of an analysis on known data training set, smart model can be signified and derived. The IF-THEN rule of classification, decision trees, and also neural network which is more advance idea. Further classification has been explained by using an example as below.



*Figure 2*: Classification Example

In this figure we can see how types of computer has been classified as into different categories and how they are further explained under each category like on the basis of size, data, brands, purpose.

## ***Decision Tree***

Decision tree has been one of the method used under classification, which has been one of the method used of getting prediction of game also it will support for getting a suitable outcome. It is one of the learning algorithm which is supervised and similar to flow diagram which include various nodes sequence, where sample value is helpful for making decision on following node to go.



Figure 3: Decision Tree Example

The figure shown above is one of the example of decision tree where we can clearly see how the tree is built it starts from is it raining which get then two roots whether is it hot or is it windy and further there is roots into bad or good. So in the starting one tree has been constructed using training data. In the second stage the use of train data helps to get prediction when require. Using the example of this tree which have a data point [“is raining”, “very windy”] it can be classed under “bad weather”. Decision tree is one of the keen learner comparative to other method of data classification. CART (Classification and Regression) algorithm are implemented as default decision tree class by “scikit learn” package. Various approach can be used to build a decision tree, however the approach which is most common is top down partition which use successive number of questions related to unknown object characteristics. Decision tree can be used in any condition where we want also it don’t have to specific on domain.

## ***Random Forest***

Complex functions can be learnt by one decision tree. But in many means it is quite prone to overfitting. Overfitting is like learning the rule which only work for a particular training set. Limit on number of rules which it learns can be adjusted in one of the method. In an example we limit the depth to 3 layers for a tree. And this tree, learns the best splitting rules of dataset, however it doesn’t learn the rules which are highly precise and which create highly exact groups by separating the dataset. This make the tree poor on overall performance, but results in trees which can have good generalization.

Many trees will be created to reimburse the performance and ask for predicting the value of class. Majority vote can be considered as overall prediction. This is the principle on which random forest works. The two major problem occur during above mentioned scenario. Firstly, it largely deterministic to construct decision tree and every time same input result in similar output. If one tries to construct multiple trees but the input will be same in every case because it has only one training dataset

## ***Advantages***

The advantages of using decision tree and random forest methods are mentioned as below

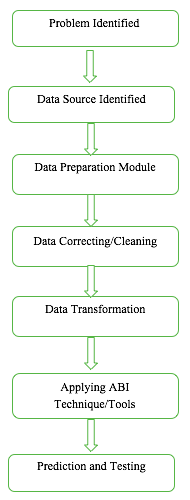
* It helps in selecting features or indirectly performing variable screening
* For preparing data, it needs small effort from user comparatively
* Relationships between parameter is nonlinear because of which performance is not affected
* It’s really use for analytics, easily interpreted, easy for explaining executives
* On large databases it runs efficiently
* Without any variable deletion it controls thousand input
* Classifiers produced are highly accurate, for most of the datasets

## ***Limitations***

* Low Execution Speed – This is very significant or important for Multi-Layer Perceptrons, in which once if the model is trained, we need to retrain the entire model and redraw the tree again if there are any dynamic updates in the dataset. (Cross Validated, 2010)
* Poor Resolution with Complex Data Relationships - Only two possibilities are possible with left or right if we use Decision Trees. This does not work if there are more variable relationships like in Fuzzy Logic. (Cross Validated, 2010)
* Subtle to small perturbations - Small changes in the dataset can affect the output of Decision Trees drastically, which means they are very sensitive to slight changes or perturbations. (Cross Validated, 2010)
* Over Fit - Random Forest technique can easily overfit for the noisy classification data and regression tasks.

# SOLUTION DESIGN AND DEVELOPMENT

In this assignment, we have tried to predict the Winning Teams in NBA using the ABI techniques and tools. The initial step to start was to make the dataset by gathering the information from the NBA matches. Below are the steps, used while finding the solution to our Business Problem.



*Figure 4*: Flow Chart of Design Solution

## ***Data Preparation Module***

Data preparation is the main step to have better prediction. The source was first identified and authenticated by checking and comparing the results with other sites. We have prepared NBA 2013-14 match data history and team rankings of 2014 as our training model and data was extracted from the website - <http://www.basketball-reference.com> which contains a lot of resources and statistical data from NBA. The dataset of 1,230 games were formed based on the extraction of information of scores and standings or rankings of teams. The raw data was extracted and stored in Microsoft Excel in which the columns act as the defining attributes and rows were utilized for storing the different information related to the games and teams. It includes names of home teams, visitor teams, points and other related data.



Figure 5: Data Source

As data preparation is very significant for making correct prediction, so high consideration and efforts were made to recognize dataset efficiency (Antonie, Zaiane, Coman, 2001).

## ***Data Cleaning/Correcting***

This step requires the data to be thoroughly cleaned and prepared for dataset. The raw data might contain incorrect or unwanted data which can affect the output or results of the prediction. So, to get the optimized or accurate results, the incorrect or erroneous data must be removed from the dataset. This unwanted data is called the noisy data. There can be different problems associated with incorrect data, redundant data and missing data.

In our raw data, there were a few columns and other information which were not required for the prediction. The date column was just a string and not a date object. The first row was blank and few headings were not complete and correct. Data also contained some duplicates including names of teams and other non-required data. These were removed and deleted using different formulae and editing tools of Excel and only the required and useful data was kept and used for dataset.

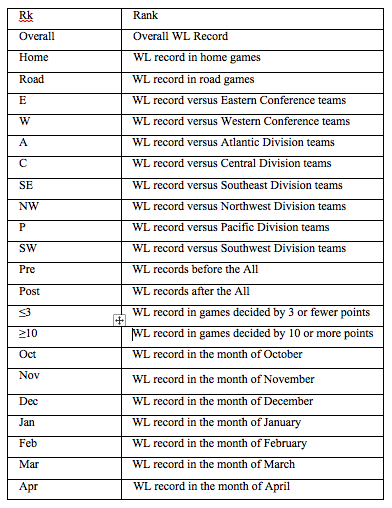
## ***3.3 Data Transformation***

The main concern was whether we have the correct or right data for prediction. The data also had null or empty fields which were then replaced with “NaN” in the dataset. There were a few teams whose names were changed in the next season. So those names were replaced with their new names in the dataset.

The data left after Data Cleaning and removing the noisy data was classified with different columns and attributes including –

|  |  |
| --- | --- |
| **Columns in the dataset** | **Description** |
| “Date” | Date of Match |
| “Score Type” | Scoring Method |
| “Visitor Team” | Name of the visitor team |
| “VisitorPts” | Points scored by the visitor team |
| “Home Team” | Name of the home team |
| “HomePts” | Points scored by the home team |
| “OT?” | Over time |
| “Notes” | Any comments or notes of match |

*Figure 6*: Columns/Attributes of NBA Season Dataset

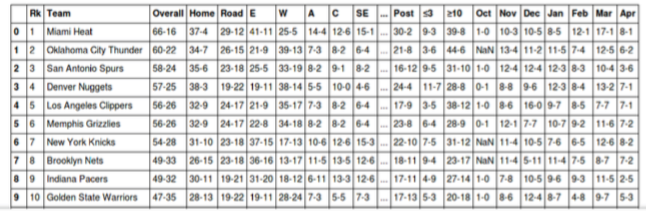


*Figure 7*: Columns/Attributes of Standings Dataset

Below are the figures showing the column names and a few data records from both the datasets –



*Figure 8*: Dataset showing column names and records of NBA Season



*Figure 9*: Dataset showing column names and records of NBA Team Standings

## ***3.4 Training of Model***

Now we have the dataset ready for the prediction and we need to train the model for getting prediction values. In this, we have extracted a few new features which we will help in improving the accuracy of the model in prediction. Below we have explained the new features which we have extracted from the dataset and used to train our model.

**Feature Selection**

We have extracted some new features which have these advantages –

* **Reducing complexity** – Having more number of features help in reducing complexities of the model.
* **Reducing noise** – Having extra new and appropriate features help in reducing chances of random correlations that do not have any meaning.
* **Creating readable models** – Having thousands of features may increase the time and difficulties to interpret for a human, so it may be worth using fewer features with appropriate meanings.

In our dataset, we have now extracted features by combining and comparing the existing data. We have to specify the class value so that the classification algorithm can compare whether the prediction is correct or not. So, if the home team wins, we assign the value to class as 1 and if visitor team wins, it is assigned value as 0. This can be computed as –

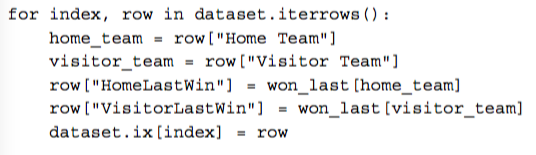


As it is a commonly proved feature that Home Team is most likely to win the most of the games in sports, so we calculated the home team winning percentage using the results as 58% of matches.

The next feature we have extracted is to check which team won their previous game, so that we can have a rough idea that whether the team has been playing well or not. We can find out this by iterating through rows and listing which team won. So, we created the default dictionary to save the team’s result as –



We can now reiterate between the rows and update the current row filling the team’s last result as below –



So we have taken two features – HomeLastWin and VisitorLastWin, to save the values of team winning their last game.

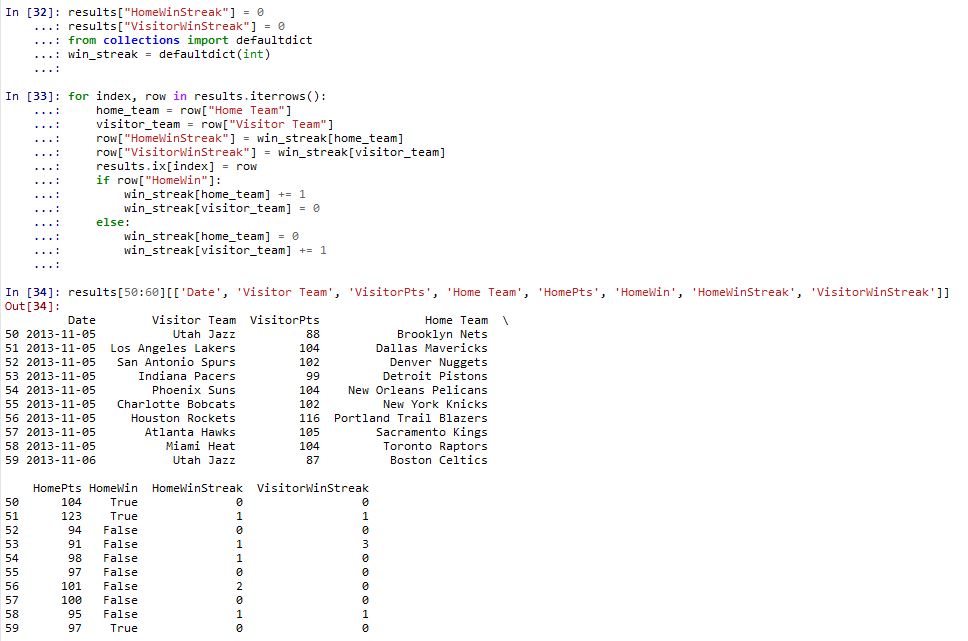


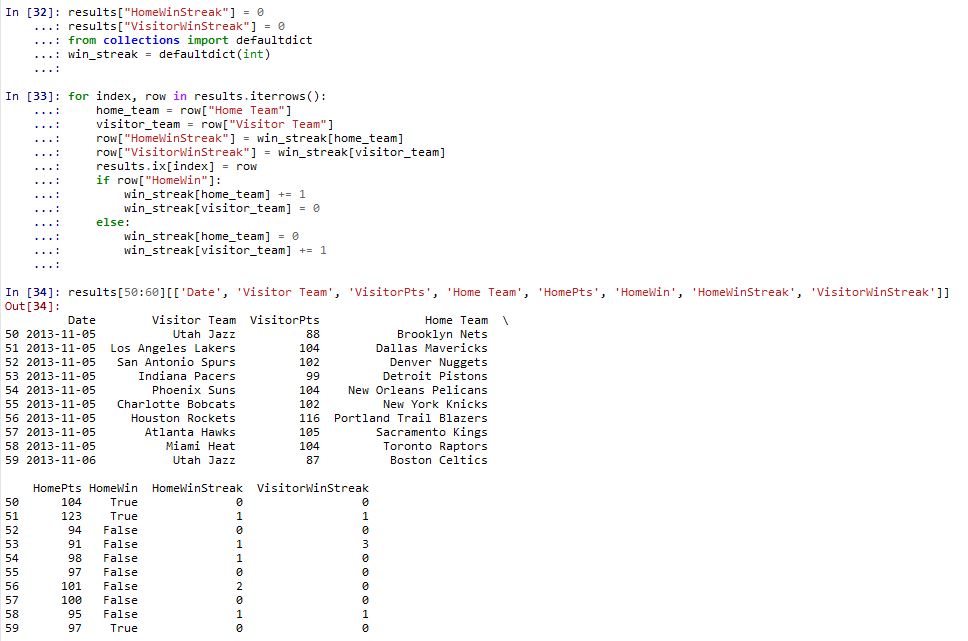
Below is the dataset with the above features.



*Figure 10*: Dataset with extra features – HomeLastWin and VisitorLastWin

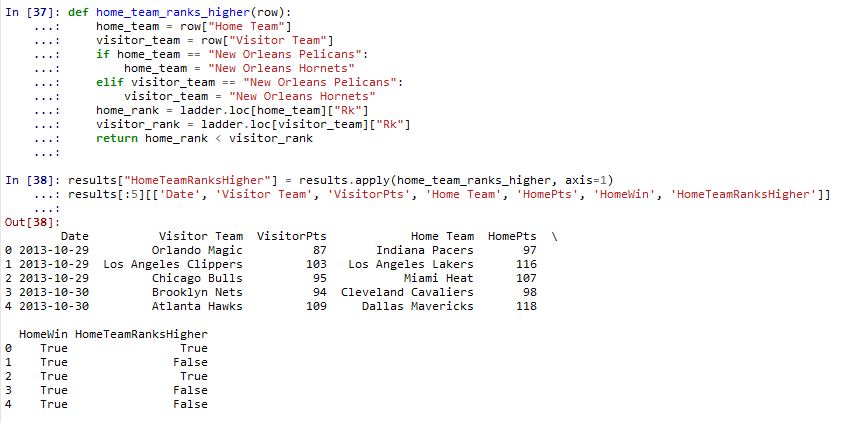
The next feature we have extracted from the data is Winning Streak. This feature not just checks whether the team has won their last game but also checks whether the team is winning their last games or not. So we have used the features as HomeWinStreak and VisitorWinStreak. These will give us an idea of consistency of team performance.

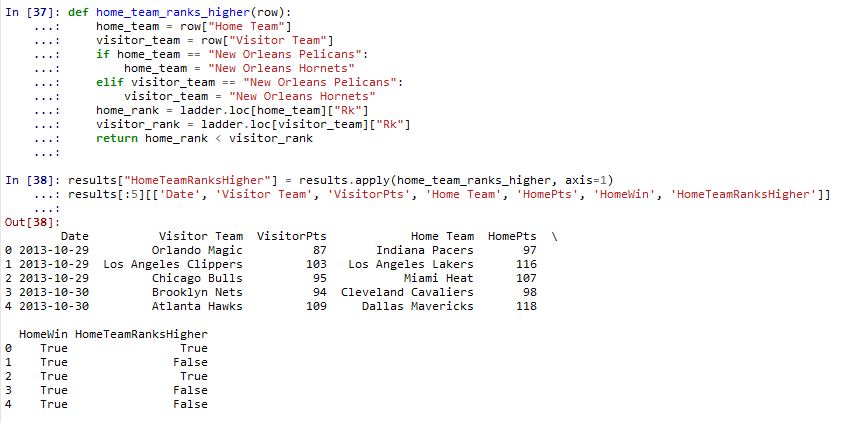




*Figure 11*: Dataset with HomeWinStreak and VisitorWinStreak Features

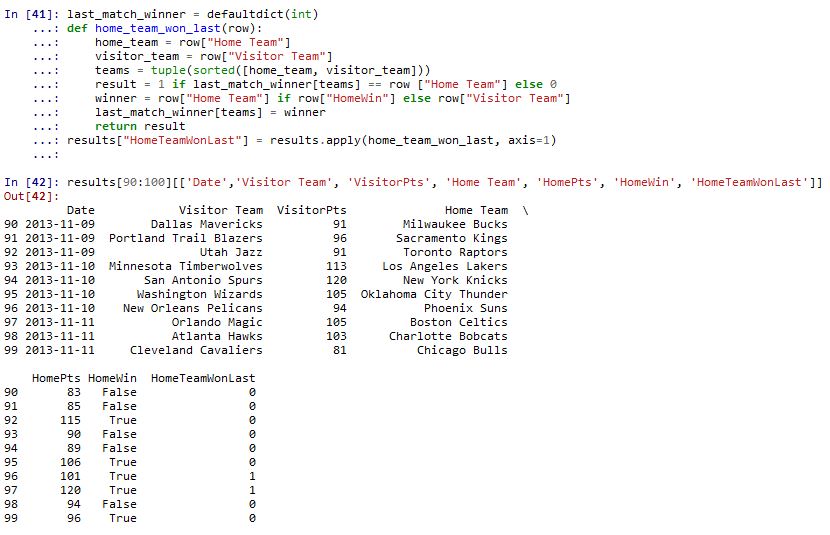
The other feature we have used from our dataset is HomeTeamRanksHigher. This feature helps in getting the information that which Home team is better than the Visitor team on the basis of points and matches won.

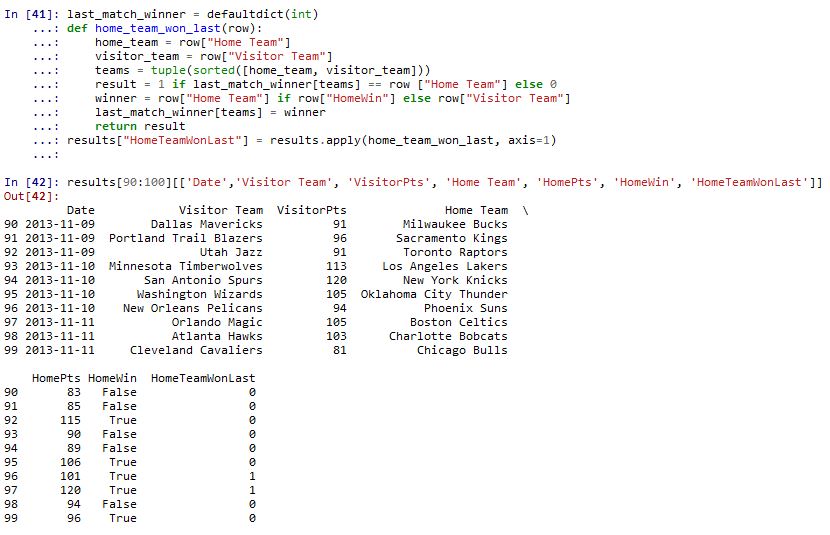




*Figure 12*: Dataset with HomeTeamRanksHigher Feature

The last feature we have used and computed is HomeTeamWonLast to check whether the home team won the last game between the two teams. So this feature looks at specifically which team won the last game not just when did they win their last game.





*Figure 13*: Dataset with HomeTeamWonLast feature

***3.5 Software Tools***

**IPython**

IPython project facilitates enhanced interactive environment which includes support related to data virtualization and also provides support for distributed and parallel computation (Fernando, Brian, 2007). IPython provides interactive command shell for multiple programming languages. . IPython is similar to eclipse IDE for java. It offers various rich media, extended shell syntax, rich history, enhanced introspection and rich history. There are following features provided by the IPython tool

* Terminal and Qt based interactive shell
* Web based Notebook for supporting inline plots, rich media, text, mathematical expressions and code
* GUI interface and data virtualization tools
* High performance and easy to use tools for supporting parallel computing

**Pandas**

Pandasis a python library that is extensively used for analysis and manipulation of the data (Layton, 2015). It is used for loading, manipulating and managing data. When performing data analysis experiments there are scenarios where same functions are implemented repeatedly for instance reading files and extraction of new features. Every time when reimplementation occurs it causes risk of bugs. Using Pandas library reduces the overhead of reimplementation of these functions and also provides confidence in using better tested code.

**Scikit-learn**

Scikit-learn is a python library that offers a range of algorithms for supervised and unsupervised learning through an interface in Python. The motive of this library is to provide robustness and support required in production system. Hence it focuses on concerns related to ease of use, collaboration, documentation, code quality and performance.

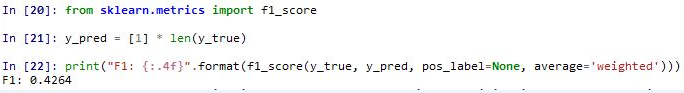
**Numpy**

Numpy is short for “Numerical Python”. It is open source module for Python that includes precompiled methods for numerical and mathematical routines. In addition it provides data structures that make efficient computation for matrices and multi-dimensional arrays.

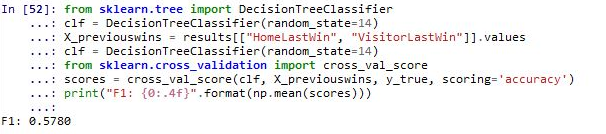
# EVALUATION

We evaluated number of features during the training of our model. There was accuracy calculation for the feature which helped in optimizing the accuracy of the model.

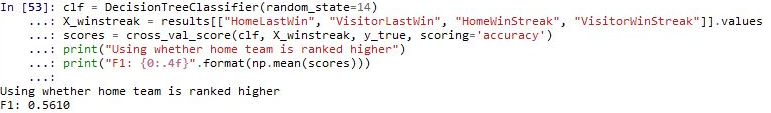
The baseline for prediction is evaluated by calculating the number of home win percentage to be 58%. This attribute “HomeWin” is used as baseline as it is true in every sport that home team wins maximum number of matches. We use f1 score to measure of test accuracy. If the value is close to 1 it is considered as best and worse if it is close to 0. The f1 score for “HomeWin” comes out to be 0.42 that is not good. Next we try to improve this score by adding new features which will improve the accuracy of the model.



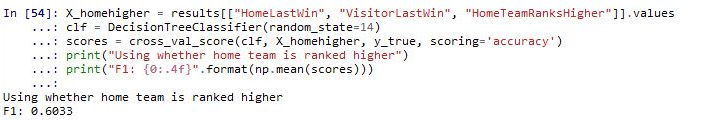
We now add two new features “HomeLastWin”, “VisitorLastWin” and using decision tree classifier to calculate the f1 score which is improved with respect to baseline f1 score.



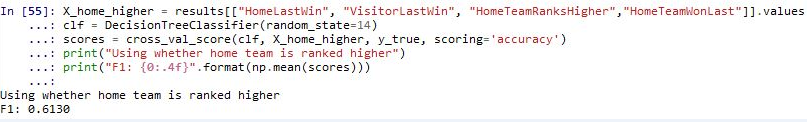
Next we add two more features “HomeWinStreak”, “VisitorWinStreak” and applying decision tree classifier again we calculate f1 score which is better than previous one.



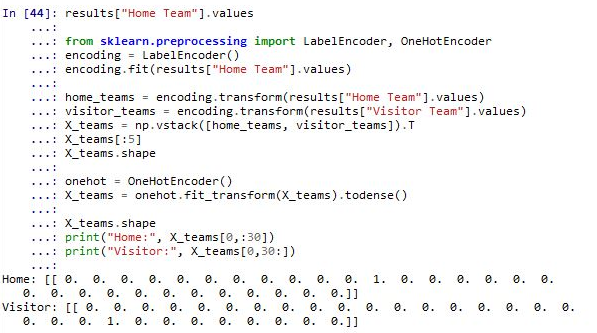
For getting the knowledge about which home team is better we use team ranking of 2013. New feature “HomeTeamRankHigher” is used which improves the f1 score.



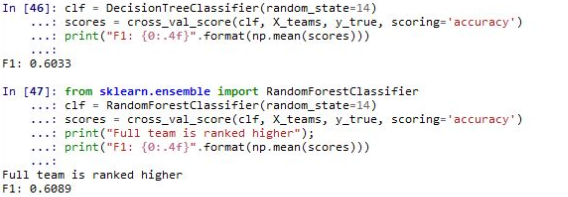
Lastly we add “HomeTeamWonLast” feature for improving the accuracy of our model.



We use one hot encoder for encoding each home team and use this feature for calculating f1 score. One hot encoder is used to convert name of each team into a combination of 0 and 1 that is essential for training of model.

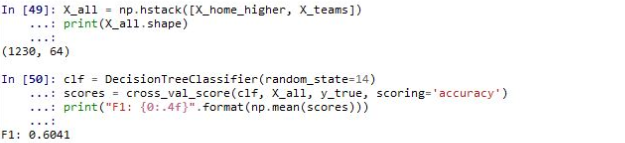


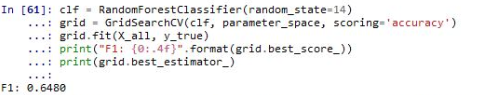
We then calculate f1 score using Teams that signifies the encoding of home teams and using decision tree and Random forest classifiers.



Random forest classifier gives better f1 score of 0.6089 than decision tree classifier that comes out to be 0.6033.

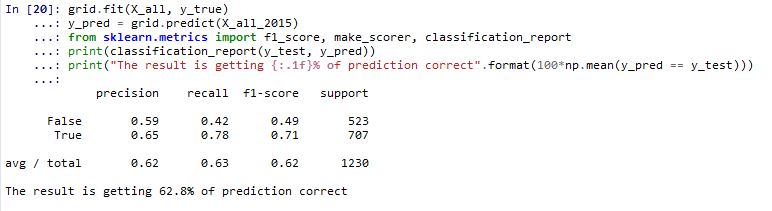
Now we combine both the team encoding feature and previous features and calculate f1 score using decision tree classifier and random forest classifier.





Random forest classifier gives better f1 score than decision tree classifier therefore we chose random forest as our prediction classifier.

Using this prediction model we predict the performance of the team in 2015 season. We have calculated the accuracy of our prediction that comes out to be 63% which is reasonably good.



# FUTURE DEVELOPMENT AND CONCLUSION

This model can be used by NBA itself or other betting companies to determine the results of the match and league as well. NBA can use this system for rating the teams, shows all the odds of all the NBA games and also showcase career graph for players. For betting companies, it can be a turnover for their revenue. Also various team managers can use it to express their strategies which can be dependent on various different parameters used in these models for prediction and to get the outcome. Also with the growth in technology there are more chances of growing this model and as more and more techniques become available, there can be more chances increasing the accuracy and reducing the errors. In addition to that marketing and endorsement companies can choose the best team of the season for promotion of their product.

On the other hands there are few factors which are uncertain and haven’t been considered for this report like competitive state, injuries, contract of players, quality of coaching, attack strength, defense strength, team strategies etc. these features affect or influence the final results to a great extent. If considered these, with other techniques the outcome can be more closely to the accurate values. Also Association rules can be generated from the decision trees which are simple and easy to understand. Association rules can be applied to the neural network which can create a better model with high accuracy and precision.

# REFERENCES

Antonie, M. L., Zaiane, O. R., & Coman, A. (2001). Application of Data Mining Techniques for Medical Image Classification. *MDM/KDD*, *2001*, 94-101.

Cao, C. (2012). Sports data mining technology used in basketball outcome prediction. *Masters Dissertation*. Dublin Institute of Technology.

Cross Validated. (2010, Aug). Retrieved from http://stats.stackexchange.com/questions/1292/what-is-the-weak-side-of-decision-trees

Fernando Pérez, Brian E. Granger, IPython: A System for Interactive Scientific Computing, Computing in Science and Engineering, vol. 9, no. 3, pp. 21-29, May/June 2007, doi:10.1109/MCSE.2007.53.

Hem Jyotsana Parashar, Singh Vijendra, Nisha Vasudeva. (2012). An Efficient Classification Approach for Data Mining. International Journal of Machine Learning and Computing, 446.

Haghighat, M., Rastegari, H., & Nourafza, N. (2013). A review of data mining techniques for result prediction in sports. *Advances in Computer Science: an International Journal*, *2*(5), 7-12.

Jermyn, P., Dixon, M., & Read, B. J. (1999). Preparing clean views of data for data mining. *ERCIM Work. on Database Res*, 1-15.

Leung, C. K., & Joseph, K. W. (2014). Sports data mining: predicting results for the college football games. *Procedia Computer Science*, *35*, 710-719.

Layton,R. (2015). Learning Data Mining with Python

Pang-Ning Tan, Michael Steinbach, Vipin Kumar. (2005).