CSE 587 – DATA INTENSIVE COMPUTING – LAB 3

Submitted By: Esther Raja Kumari Katti, Keerthana Baskaran

UB# 50288205, 50288944

Contents:

1. Title

Win prediction in sports – Cricket IPL

2. Abstract

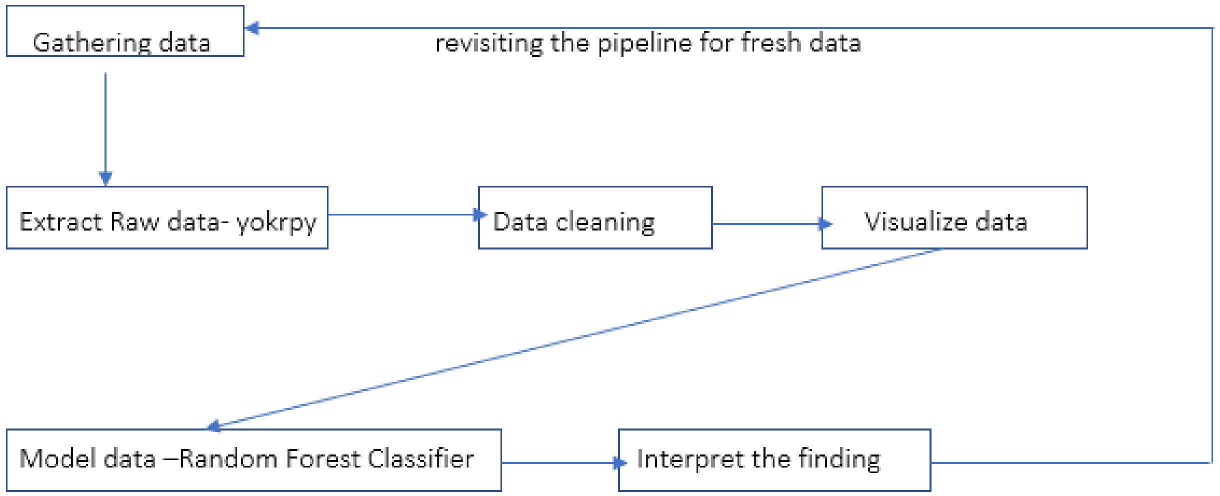
T20 has gained its importance since its introduction and has become star attraction within cricket. As a result, the analysis of a league like this IPL for win prediction is of growing importance. In this lab we are using the IPL data from 2008-2017 which is available in Cricsheet.org. We are going to use yorkr python library to extract the data from Cricsheet.org. We can use this dataset to extract key data and for statistical visualization. We are trying to predict the results of two teams accurately and predict winning team. We are going to use random forest classifier as our model.

3. Problem Statement

By solving the win prediction problem, we can predict the outcome of game and performances of team and players. We can also decide the price of player if a club wants to sell or buy. We can help connecting brands and sponsors and answer questions like which player is on downward trend, who can be traded etc.
In real time we will use huge amount of data for prediction. In cricket alone data is generated every day and historic data of 2 decades is present which can be of huge advantage for accurate and precise predictions. Using Spark for machine learning models is highly useful in increasing the accuracy and speed. Also the availability of Graphx service enables sparks to process graphs and information that is graphical in nature thereby allowing us to make predictions based on the images and not losing the image data info and increasing the accuracy.
We use Apache Spark MLib to build a model for this problem because MLib is also predominantly faster in implementation than Hadoop and is also capable of solving several problems, such as statistical reading, data sampling and premise testing, to name a few.

4. Solution (model) and design

4.1 Methodology and Pipeline architecture



Gathering data can be from internet/internal databases/third parties. Here we gather the data from Cricksheet.org. Now the raw data from internet is processed into usable format (.csv, JSON, XML etc). Here we process into csv.

Skills Required:
- Distributed Storage: Hadoops, Apache Spark.
- Retrieving Unstructured Data: text, articles, documents,yaml files.

In data cleaning part we take the information which is important to the problem. Taking the features needed, removing the duplicates. Domain level expertise is needed to discard any feature or value.

Skills Required:
- Coding language: Python, R.
- Data Modifying Tools: Python libs, Numpy, Pandas, R.
- Distributed Processing: Hadoop, Map Reduce/Spark.

During data visualization we try to find patterns and values in the data. Try to find differences and similarities through graphs, charts and analysis.

Skills Required:
- Python: NumPy, Matplotlib, Pandas, SciPy.
- R: GGplot2
- Statistics: Random sampling, Inferential.
- Data Visualization: Tableau.

After cleaning the data and finding out the features that are most important for any given problem by using relevant models as a predictive tool will enhance the decision-making process. Evaluating the machine learning algorithm is an essential part of data science pipeline. The model may give satisfying results when evaluated using a metric like accuracy score. Since we have a large dataset we would consider random forest as optimum choice for our problem

Skills Required:

- Machine Learning: Supervised/Unsupervised algorithms (Random forest)
- Machine Learning Libraries: Python (Sci-kit Learn, NumPy).

Interpreting the results at the end. Domain experts can help us in visualizing the findings and communicating them.

Skills required:

- Business domain knowledge.
- Data visualization tools: Tableau, D3.js

4.2 Dataset Description

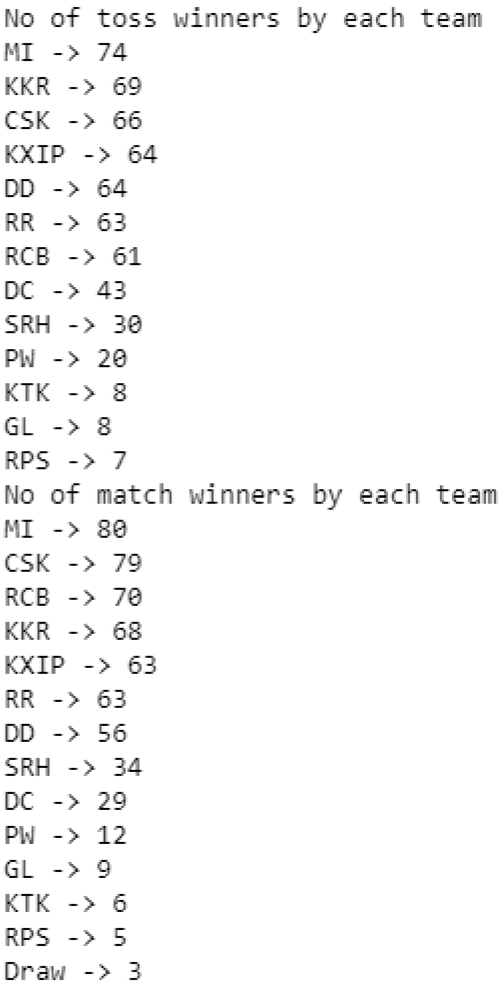
We have used the Data from *https://cricsheet.org/* which is of size 36MB. The dataset is about IPL in cricket which contains all of the matches, and the others certain sub-sets of matches, such as for type of matches, matches for certain countries, teams, or genders, or periods of time. We can extract the data from these yaml files into text or csv using yorkr package of R. This package is based on data from Cricsheet. Cricsheet has the data of ODIs, Test, Twenty20 and IPL matches as yaml files. The yorkr package provides functions to convert the yaml files to more easily R consumable entities, namely dataframes.

4.3 Solution

The data collected has to be cleaned and converted to csv. We can use BeautifulSoup for data cleaning and load the file in HDFS. Clustering the batsman and bowlers can be done using K-means algorithm with help of PySpark MLLIB. We can optimize the k value by visualizing the elbow plot for different k values. The optimized. The parameters used for clustering batsman can be Runs, Strike Rate, Average, Number of 4s,6s and 50s The parameters used for clustering bowlers can be Wickets, Economy, Average, Strike Rate
Here we followed the method of using Random Forests. The outcome of the match was predicted using the decision trees. Using the csv file generated we perform to predict the score again ball by ball. Using the Spark MLLIB which provides functions to train and construct Tree models. We can train the model, with the following parameters: Batting Average, Batting Strike Rate, Bowling Average, Bowling Economy, Bowling Strike Rate, Number of balls (represented with the overs), Innings and finally the Runs. This can be trained into the regression tree and thus is later used to predict the possible outcomes for the particular match. Thus we simulate the match give and predict on the basis of the trained model.

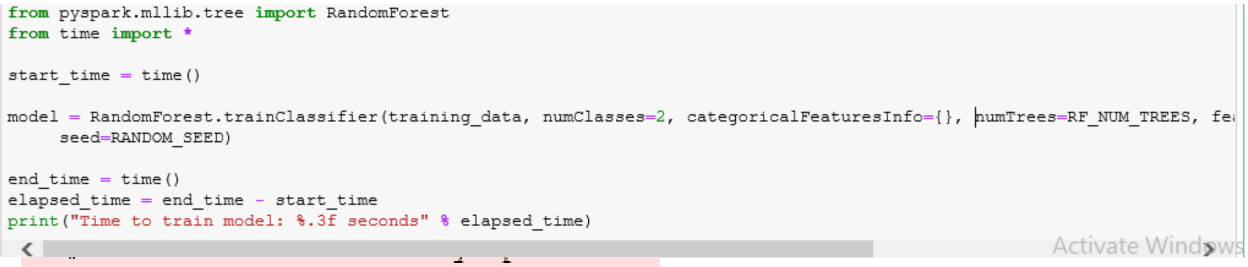
5. Outcomes and visualization

Finding if we can use Toss winner as one of the feature to decide the winner

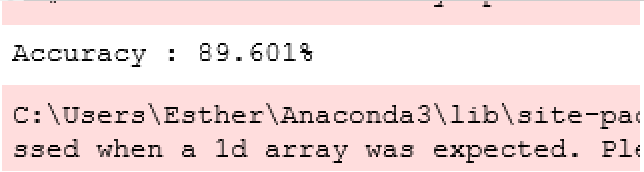


```
No of toss winners by each team
MI -> 74
KKR -> 69
CSK -> 66
KXIP -> 64
DD -> 64
RR -> 63
RCB -> 61
DC -> 43
SRH -> 30
PW -> 20
KTK -> 8
GL -> 8
RPS -> 7
No of match winners by each team
MI -> 80
CSK -> 79
RCB -> 70
KKR -> 68
KXIP -> 63
RR -> 63
DD -> 56
SRH -> 34
DC -> 29
PW -> 12
GL -> 9
KTK -> 6
RPS -> 5
Draw -> 3
```

Below is the expected accuracy and example code snippet of our model:

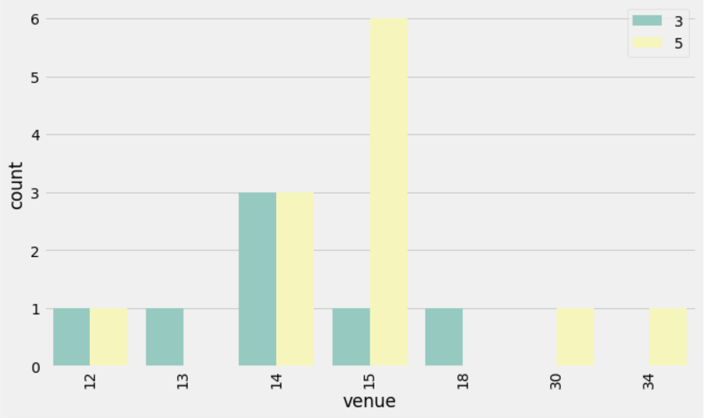


```
from pyspark.mllib.tree import RandomForest
from time import *

start_time = time()

model = RandomForest.trainClassifier(training_data, numClasses=2, categoricalFeaturesInfo={}, numTrees=RF_NUM_TREES, fe
    seed=RANDOM_SEED)

end_time = time()
elapsed_time = end_time - start_time
print("Time to train model: %.3f seconds" % elapsed_time)
```



```
Accuracy : 89.601%

C:\Users\Esther\Anaconda3\lib\site-pac
ssed when a 1d array was expected. Ple
```

Here we are analysing top 2 team based on number of matches won against each other and how venue affects them.Previously we noticed that CSK won 79, RCB won 70 matches and now comparing venue against a match between CSK and RCB we find that CSK has won most matches against RCB in MA Chidambaram Stadium, Chepauk, ChennaiRCB has not won any match with CSK in stadiums St George's Park and Wankhede Stadium, but won matches with CSK in Kingsmead, New Wanderers Stadium.It does prove that chances of CSK winning is more in Chepauk stadium when played against RCB.Proves venue is important feature in predictability.



6. Summary

- As already mentioned the amount of Data in Cricket rapidly increasing every day. Big Data plays a major role in Sports to make a prediction on Win in Sports by taking into account the past data.
- In this Lab, we have made use of machine learning algorithms, Random Forests, in order to predict the outcome of the Indian Premier League.
- We got a satisfactory result of correctly classifying 41 IPL 2018 matches correctly out of 60 total matches.
- We used the dataset from Season 1 till Season 11. These matches details were obtained after putting the dataset through cleaning pre-processing.
- Featured which we have taken into consideration for prediction purpose were Toss winning , Venue, Balling Average and Batting Average.
- The accuracy of the classifier would have improved further if the team weightage was calculated immediately after a match end.

7. References

Dataset Link:
https://cricsheet.org/
Library for extracting data:
https://gigadom.in/2016/04/02/introducing-cricket-package-yorkr-part-1-beaten-by-sheer-pace/
Example code References:
https://dzone.com/articles/ipl-cricket-analytics-and-predictive-model
Other:
https://dzone.com/articles/overview-of-the-data-science-pipeline