International Journal of Computer Applications Technology and Research
Volume 4– Issue 1, 42 - 47, 2015, ISSN:- 2319–8656

A Hybrid Prediction System for American NFL Results

Anyama Oscar Uzoma
Department of Computer Science
Faculty of Physical and Applied Sciences
University of Port Harcourt
Rivers State, Nigeria

Nwachukwu E. O.
Department of Computer Science
Faculty of Physical & Applied Sciences
University of Port Harcourt
Rivers State, Nigeria

Abstract: This research work investigates the use of machine learning algorithms (Linear Regression and K-Nearest Neighbour) for NFL games result prediction. Data mining techniques were employed on carefully created features with datasets from NFL games statistics using RapidMiner and Java programming language in the backend. High attribute weights of features were obtained from the Linear Regression Model (LR) which provides a basis for the K-Nearest Neighbour Model (KNN). The result is a hybridized model which shows that using relevant features will provide good prediction accuracy. Unique features used are: Bookmakers betting spread and players' performance metrics. The prediction accuracy of 80.65% obtained shows that the experiment is substantially better than many existing systems with accuracies of 59.4%, 60.7%, 65.05% and 67.08%. This can therefore be a reference point for future research in this area especially on employing machine learning in predictions.

Keywords: Data Mining, Hybrid System, K-Nearest Neighbour, Linear Regression, Machine Learning, National Football League

1. INTRODUCTION

Predicting the outcome of events is of interest to many, ranging from meteorologists, to statisticians, the media, to financial experts, to economists, to clubs, to merchants, to fans, to pundits and to betting markets (bookies). In the past, involvement in prediction was a leisure activity with elements of luck and experience inter operating. Now, predicting events outcome has become interesting as a research problem, in part due to its difficulty; this is because prediction outcome is dependent on many intangible or human induced factors. Predicting the outcome of event has also become a mega business but even armed with all these expertise in analyzing past data, it is very hard to predict the exact outcome of range of events.

Looking at games prediction as a sub-domain in predictions, the world has yielded huge profits and investments from these possible games outcomes. For example, NFL (National Football League) football is arguably the most popular games in North America. Over the past two decades alone, NFL has truly become America's game, with millions of people watching NFL games live on television at home and fans going to the stadium. With emerging facts that betting market in the United States accumulates nearly $1B per year on football games [1], hence, it implies that investments in forecasting outcomes in this area will be a worthy venture.

Countless number of people, computing tools and even internet websites claim they know or they can predict the possible win, lose and draw outcomes of future games. Their prediction results also come in varieties and some degree of bias, from the most accepted to the least accepted. With an avalanche of different opinions about prediction out there, the question becomes, which of these are actually predicting correctly?

With this research work, a completely data driven objective system was designed to predict the outcome of future games, purely for academic and business purposes.

2. RELATED WORK

A large number of literatures have been dedicated to the development of goal modeling, result modeling, ratings and rankings for games prediction. These works include:

[2], developed a Logistic Regression/Markov Chain Model for NCAA Basketball, in their work the National College Championships was used as their case study. Markov chain model was used for teams' ranking, the underlying model implements a chain with one state for each team. The intuition is that state transitions are like the behavior of a hypothetical voter in one of the two major polls. The current state of the voter corresponds to the team that the voter now believes to be the best. At each time step, the voter evaluates his judgment in the following way, given that he currently believes team i to be the best, he picks (at random) a game played by team i against some opponent j. With probability p, the voter moves to the state corresponding to the game's winner; with probability (1 – p), the voter moves to the losing team's state. The Logistic regression model used home-and-home conference data to estimate an answer questions from the existing problem. Good prediction accuracy was obtained with limitations on the poor ranking for losing teams and approach used only basic data.

[3], worked on A Quantitative Stock Prediction System based on Financial News. In their work the discrete stock price prediction using a synthesis of linguistic, financial and statistical techniques to create the Arizona Financial Text System (AZFinText) was done. The major objective of the project was to provide predictions for stock market using statistical data gathered from financial news. The lines of research approach used were Mean Squared Error (MSE), visualization tools and Machine Learning Techniques. Prediction accuracy of 71.2% was obtained with a Simulated Trading return of 8.50%.

[4], proposed a modified Least Squares approach incorporating home field advantage and removing the influence of margin of victory on ratings, identified key attributes of any ranking system and modeled these novel features into the new system. A prediction accuracy of 70.1 percent was obtained and the limitation of the approach is the fact that the modeled data was linear in nature.

[5], used simple regression-based technique to predict the outcome of football matches. The model investigated the linear relationship that exists between the variables and data sets. Number of games played, Scoring margin (average points scored per game minus average points, yielded per game, despite the BCS decree that margin of victory not be used for computer ratings, just to gauge the importance of scoring margin as a predictor), Offensive yardage accumulated per game, Offensive first downs per game, Defensive yardage yielded per game, Defensive first downs yielded per game, Defensive touchdowns yielded per game, Turnover margin (takeaways minus giveaways), Strength of schedule. Limitation of this system is that it is linear in nature, poor prediction accuracy of 59.4 percent.

Table 1: Summary of Related Works

AUTHOR(S)	TITLE OF RESEARCH WORK	FEATURES/ TEHNIQUES	EXPERMENT DATA SETS USED	PRED ACC	OBSERVED ADVANTAGES	OBSERVED PITFALLS
[6]	Ocean Model, Analysis and Prediction System	Root Mean Square Error	Daily height anomaly data	Good	Real time observation system	Complex approach
[7]	Advanced Regional Prediction System (ARPS)	Non-hydrostatic moel techniques Wth Perl Prog Lang	Snow assessment parameters, cloud access	Good	Real time parameters	Complex model
[8]	Prediction Model Study of Basketball Simulation Competition Result Based on Homogeneous Markov in NCAA	Markov Chain	Pre-season data	Good	Good prediction accuracy Simulation of pre-season results	Few dataset was obtained
[3]	A Quantitative Stock Prediction System based on Financial News	Mean Squared Error (MSE), Visualization tools and Machine Learning Techniques	Stock market using statistical data gathered from financial news	71.2 Percent	Simulated Trading return of 8.50% Good prediction accuracy	Complex model

3. ANALYSIS OF EXISTING SYSTEM

The existing system is the research work done by Brady T. and Madhur L. 2008. Their work involved the use of a straightforward application of linear modeling in the development of a predictive model for the outcomes of college football bowl games. The implemented techniques identifies important team-level predictors of actual bowl outcomes in 2007-2008 using real Football Bowl Subdivision (FBS) data from the completed 2004-2006 college football seasons. Given that Bowl Championship Series (BCS) ratings was used to determine the teams most eligible to play for a national championship and a playoff system for determining a national champion.

Their approach uses Linear Regression based technique to predict the outcome of football matches. The model investigated the linear relationship that exists between the variables and data sets.

3.1 Linear Regression

The existing system develops a model using a linear Multiple Regression approach which implies that more than one predictor variable is available and the linear components represents the regression coefficients being additive. The algorithm below represents the multi linear regression approach which provides a rating output.

Algorithm

Input

Attributes X1, X2…. Xn

Main process algorithm

$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \varepsilon$

$\beta 0$ = intercept

$\beta 1 \beta \rho$ = regression coefficients

ε = residual standard deviation

Output

Y= Dependent variable (Predicted Result used for rating)

Features used

The following features were used: Scoring margin (average points scored per game minus average points yielded per game, despite the BCS decree that margin of victory not be used for computer ratings, just to gauge the importance of scoring margin as a predictor, Offensive yardage accumulated per game, Offensive first downs per game, Defensive yardage yielded per game, Defensive first downs yielded per game, Defensive touchdowns yielded per game, Turnover margin (take-aways minus give-aways), Strength of schedule (as computed by Jeff Sagarin for USA Today).

Limitations of the existing system

The following limitations were observed with the existing system. Over fitting of data, Poor prediction accuracy, only linear in nature, model cannot be trained and does not provide generalization to the prediction problem.

Prediction Accuracy: A prediction accuracy of 59.4 percent was obtained.

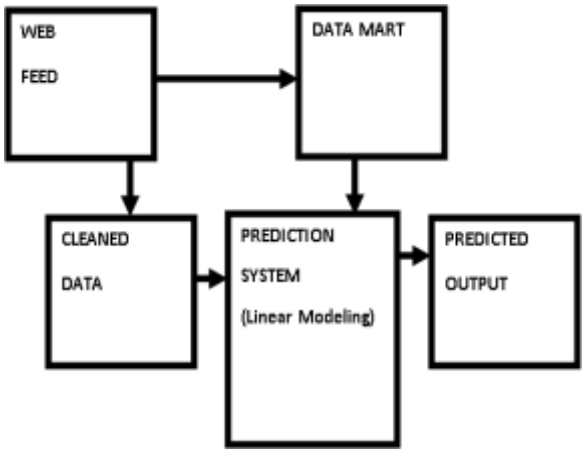

Figure 1. Brady T. and Madhur L., (2008)

(Existing System)

4. ANALYSIS OF PROPOSED SYSTEM

In the proposed system, machine learning algorithms were developed to out-perform the existing system. The proposed model framework is a hybrid of **Linear Regression Technique and K-Nearest Neighbour Technique**, which employs an objective supervised learning method. A major consideration in the choice of a hybrid system is that winning occurs in a variety of ways which in turn affects the statistics of these games. A quick glance at the various games statistics does not correctly provide the winner of a game, although it does lend some insight. In fact, there are countless examples of games in which statistics favor a team to win a game but the team eventually lost that game. Hence this indicates that there is no linear mapping method in which a winner can be chosen based solely on a group of statistics.

The hybrid system could be used to perform non-linear mapping based on a variety of relevant statistics, hence for this project the dataset to be considered will be available from the games portal. The importance and weight of each statistic must be determined prior to making a prediction using linear regression as this will provide appropriate statistical weights.

The K-Nearest Neighbour will provide classification of already weighed features. Results from trained prediction set will be applied to unseen games. The resulting hybrid model provides an optimized model which in turn will yield good results with good prediction accuracy with big implication on the various dependents of the results.

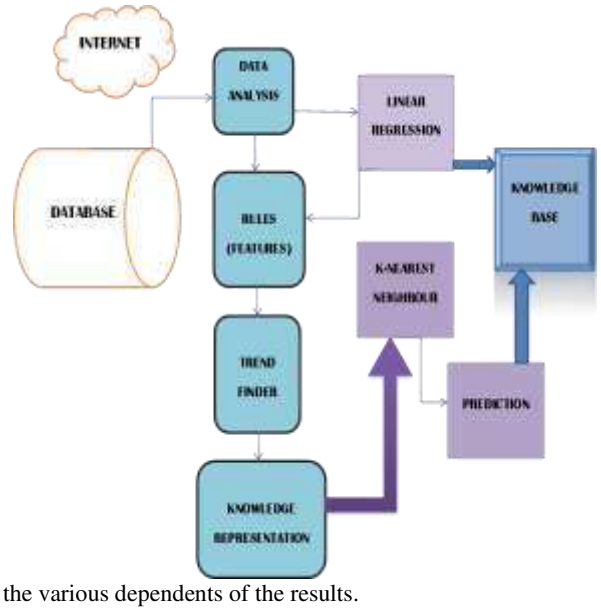

Figure 2. Proposed Hybrid Model

In designing the hybrid system, the following steps will be employed:

Step 1: Problem definition

The hybrid system will provide a correct understanding of the existing problem. Here the understanding will be broken into the project objectives and the requirements.

Step 2: Data collection and pre-processing

For the purpose of this hybrid system, there is need to acquire data from NFL box score. The dataset location on the server will be downloaded automatically from pro-football using Google's URL crawling tool and Microsoft Excel Web Extraction Feature then preprocessed with excel to acquire the right features. The dataset used is NFL games statistics for week one week 16 in the 2013 season.

Step 3: Modeling

This phase is the core of the hybrid system and will be divided into two sub steps:

i) Build model

ii) Execute model

Build

In this phase, the linear regression technique and K-Nearest Neighbour techniques will be developed using the features sets.

Execute

The resulting value of range -1 to +1 from the Linear Regression Technique will provide attributes that affects or contributes to the prediction results. The results will be used as a basis for the K-Nearest Neighbour model.

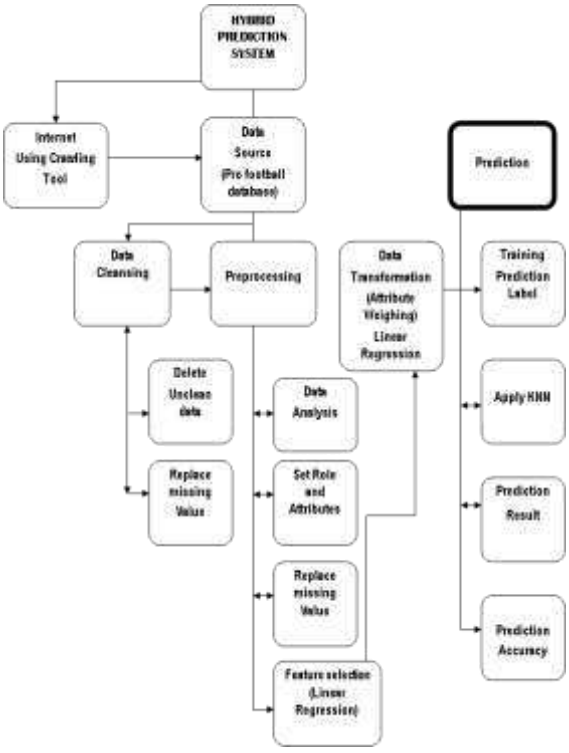

Figure 3. Algorithm of Implementation (Proposed Hybrid Model)

4.1 Features used

The following set of features will be used for the hybrid proposed system.

PtsW: Points Scored by the winning team

PtsL: Points Scored by the losing team

W#: Week number in season (Road)

YdsW: Yards Gained by the winning team

YdsL: Yards Gained by the road team

TOW: Turnovers by the winning team

TOL: Turnovers by the losing team

Tm: Points scored

Tm: Points scored

Rec: Team's record following this game (Streak)

WLD: Win, loss, draws percentage of the home team

WLD: Win, loss, draws percentage of the road team

TotYd: Total Yards Gained on Offense

TotYdL: Total Yards gained on defense

PassY : Total Yards Gained by Passing (includes lost sack yardage)

RushY: Total Rushing Yards Allowed by Defense

Sp. Tms: Special teams

Offense: Offense

Defense: Defense

Rating: Strength of team using Simple rating system

LBL: Betting Point Spread

4.2 Algorithm

The proposed model uses a linear Multiple Regression approach adopted from the existing system and K-Nearest Neighbour.

Linear Regression

Input

Attributes X1, X2…. Xn

Main process algorithm

$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \varepsilon$

β0 = intercept

β1βρ = regression coefficients

ε = residual standard deviation

Output

Y= Dependent variable

K-Nearest Neighbour

Main process

Given a query instance Xq to be classified,

Let x1, x2….xk denote the k instances from training examples that are nearest to Xq.

Return the class that represents the maximum of the k* instances.

i Associate weights with the attributes

ii Assign weights according to the relevance of attributes

iii Assign random weights

iv Calculate the classification error and adjust the weights according to the error

v Repeat till acceptable level of accuracy is reached

Standard Euclidean Distance

$d(xi,xJ) = \sqrt{(\text{For all attributes a} \sum (xi,a - xJ,a)2)}$

Output

Predicted results for weeks 16 and 17

5. RESULT

Predictions were made using both prediction sets and were tested for weeks 16 and 17 of the 2013 NFL season. In both cases, the seasonal moving average of the prediction set, proved to be very effective in predicting the outcome of the games.

The following results were obtained as displayed in the figures.

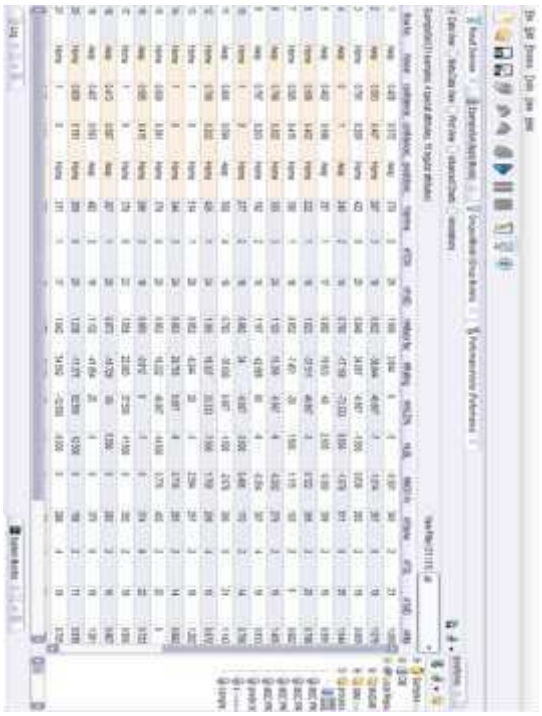

Figure 4. Predicted Results

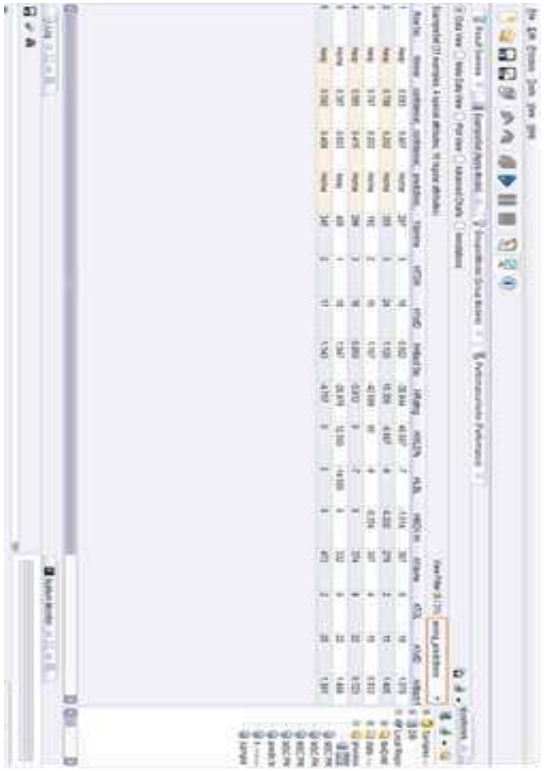

Figure 5. Wrongly Predicted Results

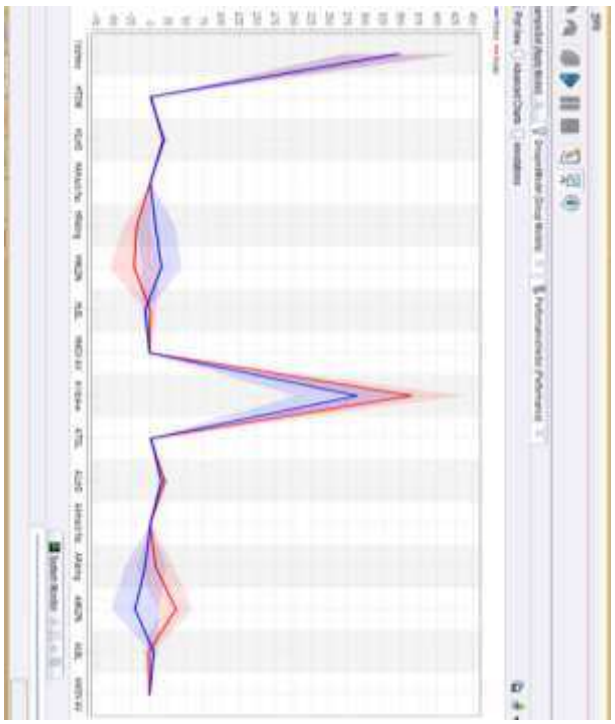

Figure 5. Graph showing predicted Results

6. DISCUSSION OF RESULTS

Prediction sets and were tested for weeks 16 and 17 of the 2013 NFL. Looking at the thirty one games that were predicted, the six games that were incorrectly predicted by the hybrid model over the progression of weeks 16 and 17, two games could be considered "upsets." An upset is when a team defeats a team with a substantially higher winning percentage in a particular game. Three of the remaining size games were games that were "too close to call," or games in which the winner is very difficult to determine. The last game is considered misclassification error in which the hybrid model predicted the incorrect outcome. The prediction accuracy of 80.65% obtained shows that the experiment is substantially better than many existing systems with accuracies of 59.4%, 60.7%, 65.05% and 67.08%.

7. CONCLUSION

The development of a hybrid model using Linear Regression and K- Nearest Neighbour techniques in the prediction of the results of NFL games with an improved accuracy.

The research highlights of this paper are:

- This paper proposes a better approach for sports prediction with unique features.
- The approach uses hybridized data mining techniques.
- The hybridized techniques used are Linear Regression and K- Nearest Neigbour.
- The results show improved prediction accuracy of 80.65%

This Research work can be considered as a successful exploration of using data mining techniques for sports result prediction and it provides a good backbone for future research works.

8. REFERENCES

[1] Borghesi R. 2007. The Home Team weather advantage and biases in the NFL betting market. *Journal of Economics and Business*, 59, 340-354

[2] Paul K. and Joel S. 2006. A Logistic Regression/Markov Chain Model For NCAA Basketball. *Journal of Naval Research Logistics*, vol 53, 1-23

[3] Robert P.S. and Hsinchun C. 2010. A Quantitative Stock Prediction System based on Financial News. Retrieved14/07/2014: http://www.robschumaker.com/publications/IPM%20-%20A%20Quantitative%20Stock%20Prediction%20System%20based%20on%20Financial%20News.pdf

[4] Harville D., (2003). The Selection or Seeding of College Basketball or Football Teams for postseason Competition. Journal of the American Statistical Association, Vol 98, 17-27

[5] Brady T. and Madhur L. 2008. New Application of Linear Modeling in the Prediction of College Football Bowl Outcomes and the Development of Team Ratings. Journal of Quantitative Analysis in Sports, 1-21

[6] Brassington G.B., Freeman J., Huang X., Pugh T., Oke P.R., Sandery P.A., Taylor A., Andreu-Burillo I., Schiller A., Griffin D.A., Fiedler R., Mansbridge J., Beggs H. & Spillman C.M. 2012. Ocean Model, Analysis and Prediction System: version 2 CAWCR Technical Report. The Centre for Australian Weather and Climate Research. No. 052Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res. 3 (Mar. 2003), 1289-1305.

[7] Ming X., Donghai W., Jidong G., Keith B. and Kelvin K.D. 2003. The Advanced Regional Prediction System (ARPS), storm-scale numerical weather prediction and data assimilation. Meteorol Atmos Phys, 082, 139–170. DOI: 10.1007/s00703-001-0595-6.

[8] Yonggan W. 2013. The Prediction Model Research on Objectives and Results of Football Game Based on Regression Method. 2nd International Conference on Management Science and Industrial Engineering (MSIE), pp139-143.