Big Data Analytics based Predictive Modeling for Forecasting Match Score

Meenakshi Srivastava

assistant professor

Amity Institute of Information Technology

Amity University Uttar Pradesh, Lucknow Campus

abstract

Prediction plays a very strategic role in improving the performance of every business, as the planning for a whole lot of other activities depends on the accuracy and validity of the exercise. The field of sports is not an exception. Analysis and forecasting play an important role in this area. With predictive analytics, the stakeholders in professional sports competitions may visualize their franchise's performance, determine the key attributes for current success or failure and predict the future performances based on the current and historical values of the key attributes of their franchise. Data analytics can be applied to discover interesting patterns from a large data set and the discovered patterns can be utilized to make predictions regarding the future possible values for a given data set. Multiple regression is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The present manuscript first the basics of big data analytics based predictive modeling have been discussed. And secondly Multiple regression is applied on various key features like mean age of players, years of experience, average team salary or points per game of players in key positions for prediction of final score for baseball match.

Keywords : Multiple regression, Big Data, Big data analytics, Predictive modelling.

I.Introduction

Big data is a large multi-disciplinary field formed by the cross section of many disciplines such as statistics, computer science, and engineering. The aim of the big data paradigm is

to provide meaningful, timely and accurate insights derived upon the collection, processing, and analysis of huge swaths of data generated in the wake of the normal operations of the modern business organization. The size of data sets involved in the big data paradigm cannot be processed by normal data collection, analysis and management tools such as RDBMS. The size of such data sets ranges from terabytes to petabytes and they require specialized frameworks for processing and analysis. Big data paradigm is applied in tandem to data engineering methods to handle the processing and collection side of the data operations of an organization. The symbiotic processes of analytics and data engineering yield insights to the decision makers which help them in taking crucial, strategic decisions which have far reaching consequences for an organization's long-term performance and its ability to attain the goals of the shareholder [1][3]]. These unique insights allow an organization to closely examine its operations and methodologies and make the necessary adjustments to its business plans and provide valuable decision-making information.

Big data analytics has following broad areas of application-

- **Law enforcement and Forensics** -Law enforcement officers and legislators can you use the predictive big data analytics to process and analyze the crime related statistics to create more effective policing and law enforcement policies and keep track of the increase or decrease in crime when subject to various attributes such as location, time etc.[2,3] Forensic experts can utilize the predictive nature of big data analytics to discover new patterns in the collected evidence [5]. Big data analytical techniques can be used to create automated detection systems for theft and fraud by analyzing the current trends to detect anomalies and trigger warnings [4] [6].
- **Risk Management** - Big data analytics helps in prediction of market trends in the future which can help firms mitigate risk. Risk management involves detection and management of risk in various fields like credit lending, risk associated with entering a new market segment. Risk modeling allows the decision makers to weigh all the options and choose the one with the lowest risk associated with it.
- **Governance and administration** - Bigdata analytics allows the government to monitor the effectiveness and the results of the policies created by it and qualify the results based on various factors. Analyzing of data related to performance of the bureaucracy of government offices can allow faster and more efficient public service institutions to be put in place which provide faster service to the government and lower the costs of operating for the government [7].

- **Marketing and advertising** -Analysis of statistics related to marketing, whether on social media, television or radio or newspaper allows companies to measure impact of their advertising efforts amongst their target demographics and find out new and unique methods of reaching their intended audience. Big data analytics also allows the firms to research the popularity of their products among certain demographics/age groups/gender and adjust their marketing strategy accordingly. On social media, an individual's interests and browsing habits allow companies to gain insights which then allow them to market their products based on an individual's social media activity.
- **Manufacturing and industry** - Firms involved in assembly line, automated manufacturing can analyze the results of their manufacturing process at each intermediate step to monitor the occurrence of errors and malfunctions in their manufacturing processes. This analysis can predict future errors and help the firm in figuring out the causes behind said errors and malfunctions in the corresponding machinery.[8]
- **Environment protection** - Pollutant emission statistics are analyzed and mined which can be helpful in optimizing energy uses of businesses and homes and further reducing their carbon footprint to lower the negative impact that they may have on their immediate environment. Big data analytics can be used to analyzes environment datasets to detect unknown containments.[9]

II BACKGROUND AND MOTIVATION

Sporting events generate massive public interest and similarly large revenues for all the stakeholders involved in the entire process of organizing, financing, and bringing the event to the intended audience. A very important part of the financial and power ecosystem of any sports competition is the "back office" or the administration, coaching, scouting and selection staff of a professional sports franchise. The people who are tasked with the responsibility of finding the right players for a sports franchise, coaching them according to their strengths and weaknesses and developing optimum strategies which can tilt the result in their team's favour. With neck and neck competition increasingly prevalent in professional sports regardless of the nature of the sports, the teams need every edge they can get to ensure their team performs to its full potential on game day. This is where the intersection of computer science and professional sports league occurs. Since mid-2000's Sports franchises in the NBA(National Basketball Association) and the MLB(Major league Baseball) in North America started to integrate statistical modeling and data analytics-based methods into their administrative structure to gain insights into

the performance of their teams and find actionable information to improve the odds of their team's victory in their respective sports competition [10].

III LITERATURE REVIEW

Bunker, Rory &Thabtah, Fadi[1] describe the application of artificial neural networks in building a classification model based on Machine learning framework which will be able to classify the outcome into one of three classes-win, lose or draw. Further, the features associated with a sports matches were divided into two categories by the authors-'external features' which are not related to events happening inside the match and 'Match-related features' which relate to the actual events occurring in the match.

Konstantinos Apostolou et al[2] analyze individual athlete performances in the sport of football based on previous season data to predict the goal scoring ability of each individual athlete and their suitable position on the field **.**

M. Manoj et al [3] applied Analytical Hierarchy Processing to predict the winner of 2017 American League Baseball championship based on the previous match statistics. They isolated 4 key factors-whether a team was playing a home game or away game, Time of day during which the match takes place(Day/Night),Ranking of the team and the division to which the team belongs. Instead of relying on finding the impact of each individual performance on the overall team performance, the above-mentioned factors were identified to be crucial to team performance.

IV METHODOLOGY

- **Feature Selection** :

The features/variables we consider having the most significant impact on performance may depend on sport to sport, but we can classify our approach to selection of crucial features based on two distinct assumptions-

A) *Team Performance forecast based on an aggregate of Individual player performances and other independent features* -

In this approach we assume that a team's final performance (win, lose or draw) is a variable dependent upon the cumulative effect of all the individual performances on the team and only including factors destined to influence performance such as quality of players, home/away games etc. This approach considers the contribution of each individual player to have a direct impact on the result and the features selected for analysis

are related to the individual performance metrics of each player related to their sport. In sports competition where a hierarchy of special roles may exist on the field such as a captain(Cricket), forward striker(Soccer), guard defense(Basketball), wicketkeeper(cricket) etc. we may choose to assign special values called "weights" to each player's performance based upon the importance of their role to their team such that a player with more important role in the team will have a larger contribution on the end result variable in comparison to a player with a less important role [7] [10] [15]. This measure allows the team performance variable from being skewed due to changes in the performance of less important team roles.

B) Team performance forecast based on the overall team performance in the past -

In this approach we consider the forecast of a team's performance variable in the future solely dependent on its performance in the past ie, an aggregate of team's wins, losses and draws over the course of a particular period under consideration. For this approach, the features we are interested in as input to our prediction algorithm would be limited to win/loss ratios in each period. This approach is much more simplistic as it reduces complexity by reducing the consideration of a myriad of individual variables impact on team performance although there is a trade-off for this simplicity as this approach is too rooted in historical trends and fails to account for subtle changes in the current team roster or strategies. After collecting and pre-processing data based on either of the above assumptions, we must apply a modeling algorithm on the data set.

- **Overview of Big Data Cycle** :

The application of big data life cycle to any project requires the aid of data engineering tools for the performing of collection, organization, and processing on the collected data set. The big data life cycle for any project begins by determining the problem statement, which in our case is the analysis of sports completion data to predict the chances of winning. Next, we look for a reliable repository of data which stores accurate and pertinent information in a structured manner for any sport under our purview. Such data can be found on the websites of regulatory bodies of a sport or professional organizations or leagues. Further, we select a subset of all the available data which will be relevant to our analytical model. All selected data is then gathered in a data warehouse or data mart. This will be followed by data cleaning to improve the accuracy of data and data transformation to organize the collected, cleaned data. Further, an analytical model appropriate for the gathered data is created depending upon the needs of the project. The Analytics phase of the big data cycle, which is also known as Data mining/Knowledge

generation, involves the application of supervised or unsupervised learning algorithms to a data set to generate unique insights. Many different types of Analytical approaches can be taken to a problem depending upon its nature, for our problem statement and domain we will utilize Predictive analytics, a type of data analytics which utilizes past and current data sets to make forecasts and predictions about future events and spot trends for the future based upon current scenarios. In the last step of the big data life cycle, the accuracy and usefulness of the analytical model is evaluated by studying the patterns revealed by the model. The timeliness and uniqueness of the insights bore by the model decide the ultimate usefulness of the analytical model. The following steps outline the process of applying the big data paradigm for solving a given problem statement:

1. *Data preprocessing*

Data preprocessing is the most preliminary stage of the big data cycle which involves the careful selection of the subset data from a larger data source, and it is cleaning and transformation for the purposes of standardization or normalization, according to the nature of the collected data.

1.1 *Data selection and organization* : Data is carefully selected to ensure that it is relevant to the data mining problem. Selecting more pertinent data reduces the volume of data and allows for quicker preprocessing. In the context of this problem, the data to be collected is found on websites storing records of statistical data generated for any sport in a semi-structured, structured, or unstructured, plain-text format. We refine our requirements by narrowing down the vast amount of data by using qualifiers such as year of play, rule set followed, average age of players etc. and many other attributes which are unique to each individual sport that can be used to select a useful data subset. The data is fetched from a target source can be in an easily manageable format such as .csv(Comma Separated Values) or .json(JavaScript Object Notation) which allow the representation of data in attribute-value pairs or 2 Tuple representations in plain text files which can be imported or exported for data processing and analytics. If the targeted data is not provided in the structured .csv or .json formats and instead present in plain text format on a website, we can use automated web scraping through various Python libraries such lxml which can be used to parse HTML and XML web documents and the associated tags.

Data is then organized based on its inherent properties such as the type of data, source of data, available volume, and variety of data. Data organization allows us to get a bird's eye view of all the data collected and informs us of the scope and range of the data set. Data organization also improves our understanding of the problem statement as we get a

glimpse of the data set which must be analyzed to provide a solution for our given problem statement.

1.2 *Data Cleaning* :

The data collected during the data selection step, either from a structured/unstructured or semi-structured source will contain some amounts of inconsistencies and inaccurate values that will produce a bias in our analytical model if used directly without any preprocessing. Hence, we need to examine and clean out collected data to properly handle any impurity, inconsistency, or inaccurate values to make sure that our model is fed clean and consistent data for its analysis. The various steps that could be taken to clean the data depend on the type of inconsistency present in the data.

- Missing values- For missing values, the absent values may be extrapolated or interpolated using the Scipy library modules of Python or series. interpolate API from pandas. Another approach would be keeping the missing value as it is.
- Inaccurate Values - Inaccurate values can be deleted or replaced with correct values.
- Outliers' detection and treatment- Outliers are values which are very much larger or smaller than all the other observed values in the data set. Outlier detection can be performed by using graphical tools such as Histograms and Boxplot. This visual representation allows us to highlight the outliers in the data set. For outlier treatment if the observation is invalid or inaccurate it can be simply discarded. Otherwise, we can use a capping technique to impose upper and lower limits on the data set and any values beyond these values are brought back inside of these limits.

1.3 *Data standardization and normalization:*

As changes in the rule sets, scoring systems and playing styles are inevitable with the passage of time in each sport, the dataset across different years can only be properly aggregated and compared after it has been normalized and standardized to ensure comparability between data sets belonging to different years. We can choose between standardization or normalization based on the needs of our project and the nature of the data present. Standardized data has a mean of 0 and it has a standard deviation of 1. Normalization is a technique used to reset the scale of various selected features to eliminate the differences in the range of the selected data and promote easier comparison between features that have been extracted from various sources. Normalization can also reduce the overall range of a data set and generally limits the range to between 0 and 1. Normalization can be performed by finding the z-score for each individual value, which is

calculated by subtracting the mean of the division from each data value. and dividing this intermediate result by the standard deviation.

2. *Analytical Model*

The analytical model will utilize the cleaned, transformed historical data to provide a classification or prediction-based forecast for a particular sport result. The analytical model can utilize a supervised learning algorithm or an unsupervised learning algorithm to find future trends. A supervised learning algorithm is a type of learning algorithm that consumes a labeled data set, known as the training set consisting of a labeled input and a corresponding expected output. This helps the algorithm in classifying any new unseen or unknown data points which are unlabeled. Supervised learning algorithms in situations where the data set in question is labeled appropriately enough for the algorithm to learn effectively labeling newer, unknown data points. Unsupervised algorithms on the other hand do not require a preexisting labeled training set and aim to classify the inputs based on inherent similarities amongst them.

V MATCH WINNER PREDICTION USING MULTIPLE REGRESSION MODEL

As the score during a match is measured by using continuous variables, meaning quantities that can take any given value between a maximum and minimum value, we can apply multiple linear regression on the selected features to create a model for score prediction. We use a multiple regression model because there are multiple independent features that influence our dependent variable, which is the team's match score. The independent features selected should have linear relationship with the dependent variable which can be checked using a scatterplot. The closeness of a point to the score line indicates the degree of linearity between that feature and the score and hence the appropriateness of selection for regression analysis. Examples of features include Mean Age of players, Years of experience, average team salary or points per game of players in key positions etc. [13] [15].

Using Multiple Linear Regression, we aim to find a linear equation which models the relationship between these features and the team's score. We aim to find a straight line which best fits all the mapped features by minimizing the least squared differences for each feature.

Multiple regression is implemented in any Python development environment by using sklearn. linearmodel LinearRegression class from the sklearn library. Next, we load .csv

or .json file containing clean, normalized features and apply the needed functions from the sklearn library to create a prediction model [11] [16].

To find the actual winner of a match, we build score prediction models for both teams involved with their respective independent features and then compare the value of dependent variable for both teams to determine the winner. The team with the larger value will be the winner.

VI OUTCOME AND CONCLUSION:

In this paper we explored the process of applying the big data paradigm to sports competition data sets to forecast the outcome of a competitive match between two teams. This real utility of a predictive analytical model is not in the simple prediction of a winner or loser but in the fact that an analytical model allows a team's management to observe the increase and decrease in their chances of success with respect to the features utilized in a particular predictive model. Modifying each individual feature to observe the changes in chances of success allows teams to simulate the results of each approach and then fine tune their strategy and focus on aspects of their play to achieve optimal results. Teams can measure the accuracy of prediction by comparing it with real world results to find the soundest analytical model. To demonstrate, we choose to predict the future match ups of the Baseball team Los Angeles Dodgers. We choose the feature Batter's Average age for given season of play as our key feature and build a regression-based model to find the runs scored per game by our subject team depending upon the average age of the batting squad. Baseball is chosen for demonstration purposes as some s external factors such as weather/pitch/ball conditions can have less chance of affecting the outcome of a match than core features related to actual play. Scatter plot of average age of batters(independent variable) and the runs scored per game per year(dependent variable)

The data consists of the attributes Run per game(Dependent variable) and the Batter's average age(Independent variable) of the Los Angeles Dodgers over the course of 12 seasons ie, from 2008 to 2019.

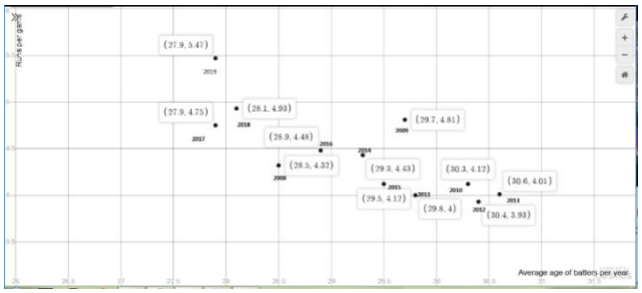

Figure 1. Average runs(Y) vs. Average age(X) and Year

This plot displays individual data points showing the average age of the team for a particular year and the average runs scored per game by the team for that year.

Function	Value
Mean of x	29.24167
Mean of y	4.4475
Correlation Coefficient r	-0.79015
a=MY - bMX = 4.45 - (-0.38*29.24) =	15.4936
b = SP/SSX = -3.95/10.47 =	-0.37775
$\hat{y} = -0.37775X + 15.4936$	

Figure 2 Generation of best fit line

Displaying the summary of the statistical measures calculated from the given data used to calculate the best fit line.

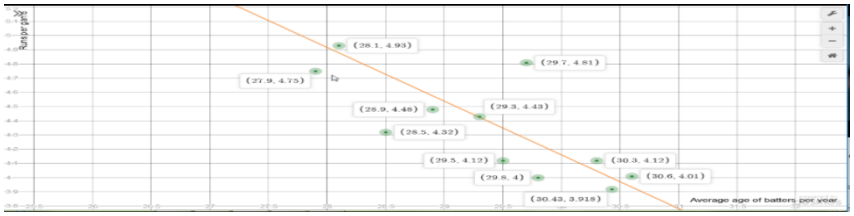

Figure 3. Best fit line scatter plot

Display of the best fit line calculated using linear regression for the selected data points using the values calculated in Figure 2s.

VII FUTURE SCOPE

For future development we can implement a classification based analytical model utilizing the support vector machines algorithm to forecast sports matches. Alternatively, a logistic regression algorithm can be used to provide a bi-variate categorical prediction model.

VIII LIMITATIONS:

1. Identification and selection of key features requires a more structured approach.
2. Lack of flexibility due to reliance on historical data.
3. Assumption of total independence in selected features in the multiple regression model while some features may be dependent to some degree.

REFERENCES:

[1] Bunker, Rory & Thabtah, Fadi. (2017). A Machine Learning Framework for Sport Result Prediction. Applied Computing and Informatics. 15. 10.1016/j.aci.2017.09.005.

[2] K. Apostolou and C. Tjortjis, "Sports Analytics algorithms for performance prediction," 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA), PATRAS, Greece, 2019, pp. 1–4, doi: 10.1109/IISA.2019.8900754.

[3] M. Manoj, R. Prashant, V. Parikh, and A. Chaudhary, "American League Baseball Championship 2017 Prediction using AHP," 2018 International Conference on Communication, Computing, and Internet of Things (IC3IoT), Chennai, India , 2018, p. 469–473, doi: 10.1109/IC3IoT.2018.8668120.

[4]McCullagh,J.,2010.DataminginsportsANeuralNetworkApproach.Intl.J.ofSciencesandEng,3,pp.131–138.
[5]Tan,PN,Steinbach,M.&Kumar,V.,.IntroductiontodatamaningPearsonAddisonWesleyBoston.[6]Landwehr,N.,Hall,M.&Frank,E.,(2005).Logisticmodeltrees.MachineLearning,59(1) ,pp.161–205.
[7]
Blundell,J.,(2009).NumericalAlgorithmsforPredictingSportsResults.UniversityofLeeds,SchoolofComputerStudies.
Zhang,S.,Pan,Q.,Zhang,H.etal.Predictionofproteinhomo-oligomertypesbypseudoaminoacidcomposition:Approachedwithimprovedfeatureextraction and Naïve BayesFeatureFusion.AminoAcids30,461–468(2006).https://doi.org/10.1007/s00726-006-0263-8
[8]Footballbetting–theglobalgamblingindustryworthbillions,"http://www.bbc.com/sport/football/24354124"
[9] NaïveBayes, http://scikit-learn.org/stable/modules/naive_bayes.html.
[10] SupportVectorMachines, http://scikit-learn.org/stable/modules/svm.html [11] HistoricalFootballResultsandBettingOddsData, “http://football-data.co.uk/data.php”.
[12] PinnaclevsMarkLawrenson, “https://www.pinnacle.com/en/betting-articles/Soccer/Mark-Lawrenson-vs-Pinnacle-Sports/VGJ296E4BSYNURUB”.
[13]LogisticRegression,"https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html".
[14]KeshtkarLangaroudi,M.,Yamaghani,M.(2019).SportsResultPredictionBasedonMachineLearningandComputationalIntelligenceApproaches:ASurvey.JournalofAdvancesinComputerEngineeringandTechnology,5(1),27-36.
[15] K.ApostolouandC.Tjortjis(2019)SportsAnalyticsalgorithmsforperformanceprediction, 10th InternationalConferenceonInformation ,Intelligence,SystemsandApplications(IISA),PATRAS,Greece,2019,pp.1-4,doi:10.1109/IISA.2019.8900754.
[16]
Thabtah,F.,Zhang,L.&Abdelhamid,N.NBAGameResultPredictionUsingFeatureAnalysisandMachineLearning.Ann.Data.Sci.6,103–116(2019).https://doi.org/10.1007/s40745-018-00189-x

[17]ZhuP.,SunF.(2020)SportsAthletes'PerformancePredictionModelBasedonMachineLearningAlgorithm.In:AbawajyJ.,ChooKK.,IslamR.,XuZ.,AtiquzzamanM.(eds)InternationalConferenceonApplicationsandTechniquesinCyberIntelligeATCI2019.ATCI2019.AdvancesinIntelligentSystemsandComputing,vol1017.Springer,Cham.https:// doi.org/10.1007/978-3-030-25128-4_62