Study of Data Behaviour and Methods for Data Prediction and Analysis

Kiranraaj Sekar[1], Amit Kumar Tyagi[1,2][0000-0003-2657-8700]

[1]School of Computer Science and Engineering, Vellore Institute of Technology, Chennai Campus, Chennai, 600127, Tamilnadu, India.
[2]Centre for Advanced Data Science, Vellore Institute of Technology, Chennai, 600127, Tamilnadu, India
kiranraaj.shanmuga2020@vitstudent.ac.in, amitkrtyagi025@gmail.com

Abstract: The age-old tricks of predicting based on gut intuition have been overruled by mathematics since day one. The study of dispersed data and being able to predict how the data is spread around the median only to show us how other similar data could be spread could also be used to predict the behaviour of data. And such is this research paper's goal to discuss the methods and help as we analyse and build models. Building regression models, analysing weather forecast models, pondering over time series models will be the goal of this research paper. At the end of the paper, one should be able to understand how data can be studied and used to build a predictor of their own, to guess according to the given specific situation. Building data models is essential in data analysis and prediction as its the core building block as to how the information is interlinked. We will use the IPL Match data's to inspect the data flow and foresee the future as an example to build a .py file.

Keywords. Data Prediction, Machine Learning, Automate Analytics, Regression Models.

1.Introduction

There is an ever-increasing demand for STEM related fields due to the recent trend of evergrowing technological requirements and demands. Education in computer fields has increased ever so in the last 20 years more than ever. And also, there has been an increase in Indias GDP with almost 9% due to the ever-expanding IT sector. In today's world Data has become as essential as breathing. Information enhances the assosciations ability to find out the issues more efficiently. They tend to point out their weaknesses in a statistical manner. With data analytics coming into picture, we can predict the outcomes in the near future and if its disadvantageous to the assosciation they can make an effort to change how the things work. But nowadays we just don't hear about data. Very commonly heard term is big data. Big data is an umbrella term alluding to enormous volumes of organized and unstructured information that we gather and produce consistently. Each client/brand collaboration is a wellspring of information, similar to the substance that organizations distribute on their site, and the activities that web crawlers record every day. Here is a comparison between different units of data:

Big data in lay man's term sounds like it's a huge incomparable data which cannot be comprehended by normal computing devices. Well in actuality it can be stored and read by a highly powered registering gadget. The 'Big' in big data is subjective to

the current technology and amount of data present. For example, a few decades ago 1 GB would have been viewed as big data as there was only 1.5 EB of data. But in todays date 1GB is not Big data compared to zettabytes of information available around the world, and it's a good idea to discuss huge information beginning with something like 1 terabyte. If we somehow happened to place that in more numerical terms, then, at that point, it appears to be normal to discuss Big Data with respect to datasets which surpass absolute information made on the planet partitioned by 1000^3.

When users think about it this way, as data is money, then the number of jobs in data analysis started to go through the roof. Information Science and Data Analytics are two of the most common words of the year. Today, information is more important to businesses than oil is to the people who work for them. In the beginning, information is put together in a simple way by an organization. This information is then used for a different purpose. This cycle helps the business fill in the gaps. However, who will do this work? In this case, who will be in charge of the information? so on Everything is done by a Data Analytics and a Data Scientist, who work together. Normal job growth in the U.S. is 3.7 percent for all jobs. There's more evidence that there's been a rise in the number of people who want information analysis experts, especially among leaders and people in power. IBM additionally predicts that interest for information driven leaders will increment by 110,000 of every 2020.

Predictive analysis is a study of data and its behavior and thus identifying how the data moves according to time. This makes it possible to create future foresights with the help of past available data. The past data can be used to make a guess as to how the data flows according to time. We can build models around the existing data and choose the one that has most accuracy. Accuracy is important in data analysis as no one would want a model with less accuracy to be making foresights. This will become important to the future associations as it will bring them lots of fortunes and also help them save a lot.

Data analysis is derived from a lot of already existing technologies like data mining, AI and machine learning. As all of these technologies are related to data, they are interlinked and bound to have common concepts and ideologies of working. This is very essential in terms of big companies and associations as they can save up a lot of many avoiding the errors and mistakes if they were able to know the future output. Each company has a different requirement for data and hence can build a model appropriate to filter their data. This can not only help the associations save but buy in their requirements and make purchases knowing the approximate results. There bound to be a data analyst team in a bigger association due to these reasons as it is very efficient and cost saving. Data analysis can be used to figure out misconducts and faults before they cause any serious harm to the company. They can also be used to check the clients' behaviours and warn us beforehand if there is any strange behaviour thus saving from company damage like cyberattacks. Associations today utilize prescient investigation in a practically perpetual number of ways. The innovation helps adopters in fields as various as money, medical services, retailing, neighborliness, drugs, auto, aviation and assembling.

Following can be a use of data analysis:

For web-based companies then can use the login details to keep a track of the number of users they have been getting at a particular time and hence able to identify the events that happened during that time which helps them figure out the user needs. *Following are the few questions the research paper will answer:*

- Can every data be predicted as to what will the outcome be, in the future, ie is it necessary for all kinds of data to be predictable?
- How should one go forth with establishing a prediction model?

2. Related Work

This includes ANOVA, linear regression (ordinary least squares) and many other types of predictive modelling. These include ridge regression and time series. Decision trees and neural networks are other types of predictive modelling that can be used. Choosing the wrong modelling method can lead to incorrect predictions and residual plots that don't have the same variance and/or mean.

2.1) Regression Analysis

Regression shows how features are linked together in terms of how they work. Linear regression models can be quickly made from covariances, but they only work with linear relationships [1]. Regression analysis is used to predict an infinite target variable from one or more independent variables. Regression analysis is used to look at natural variables, not variables that change through experimentation, which is what it does. As we said earlier, there are many different types of regression. Once we've decided to use multivariate analysis, how can we choose which regression method to use?

2.1.1) Anova

When the final variables are continuous and the input variables are both categorical and discontinuous, Anova comes into play to help figure out how to get the best results. Null hypotheses think that all groups are the same and that the data should be spread out about the same, which means that every other sample case will have about the same variance.

2.1.2) Linear Regression

Linear regression model is considered when:

Resultant variable = Summation ($K_i * X_i$) I ranging from 0 to n, n = Number of variables.

In non-mathematical terms, the resultant variable is linearly dependent on the input variables. The input variables can be continuous or discontinuous meaning categorical.

2.1.3) Logistic Regression

Logistic regression [2] is used primarily with dichotomous dependent variables, the technique can be extended to situations involving outcome variables with 3 or more categories (polytomous, or multinomial, dependent variables) / give an overview of the logistic regression model / discuss the main similarities and differences between logistic regression and linear regression and the basic assumptions of logistic regression / use data from a hypothetical study to show how to interpret a logistic regression analysis / in particular, [the author reviews] how to interpret model coefficients, test hypotheses, and interpret classification results / use data from actual research studies to show how to interpret logistic regression analyses that involve more than 1 predictor variable / describe model-building procedures for studies

that have many potential predictor variables.

2.1.4) Time series

Multivariate time series analysis can be used to predict future responses based on how people have responded in the past. In order to make a statistic, you need to know about how a variable changes over time. The information is bivariate, so the experimental variable is time, and so is the information. The rows must stay the same over a long period of time; H. are normal: the mean value and variance of the series stay the same over a long period of time. Another thing: Even if the residuals aren't correlated, they still need to be normal even if they're spread out over time. Outliers must not be in the series. As it turns out, when there are random shocks, they should be random. They should be random, with a mean of 0 and a long-term variance.

2.1.5) Ridge Regression

There is a lot of talk about how to use biased estimation in data analysis and model building. For ridge regression, researchers look at how it works and how it relates to generalised inverse regression. Researchers also show the results of a simulation experiment and three real-world examples of how ridge regression can be used. Negative comments are made about how to choose variables, how to test one's model, and how to compute ridge and generalised inverse regressions. Ridge regression is better at predicting and extrapolating than least squares when the predictor variables are very similar. It's also a safe way to choose variables.

2.1.6) Decision Tree

Decision tree is composed of various decisions and hence the entire model is built on different paths that could lead to different results [10]. Computing decision trees can be vastly time consuming as it expands to m^n (m = no of decisions) i.e exponentially. But for data that heavily depends on decisions its suitable to go with the decision tree model to predict the results.

2.1.7) Neural Networks

There are a lot of layers in Deep Learning or Deep Neural Networks, which are Artificial Neural Networks (ANN) [4, 11]. Over the last few decades, it has been thought of as one of the most powerful tools. It has become very popular in literature because it can handle a lot of data. The desire to have more hidden layers has recently outpaced the performance of traditional methods in a number of different fields, especially pattern recognition, which is where this is happening. Many people like the Convolutional Neural Network, a type of deep neural network (CNN). A mathematical operation called "convolution" is what gives it this name. CNN have a lot of layers, like a convolutional layer, a non-linearity layer, a pooling layer, and a fully-connected layer. The convolutional and fully-connected layers have parameters, but the pooling and non-linearity layers don't have parameters, so you can't set them. The CNN is very good at machine learning problems. People who work with images, like the largest set of images that have been classified.

3. Proposed Methodology

Following are the points that were written down as a guide for making a predictive model:

I. Preparing business goals: The information or the business goals can be used with a chosen algorithm or model. Other times, the simplest way to do

something isn't so clear-cut. Compare the results. Users should choose the best model based on what other people have said. Sometimes, users prefer to run a group of models at the same time and choose the best one by comparing their outputs.

II. Preparing data: The information is usually spread out across a lot of different sources and needs to be cleaned up and prepped. Data may have a lot of duplicate records and outliers. Based on the analysis and the business goal, users decide whether to keep or get rid of them. Also, the data may not have all of the information users need, may need to be changed, and should be able to generate attributes that are more predictive for ones goal. As a general rule, the quality of the information shows how good the model is.

III. Training and test datasets will have to be split up. Users build the model with the help of the training dataset. Users use the test data set to make sure that the model's output is correct. Because if users don't, users could end up overfitting ones model, which is when users train ones model on a small set of data so that it picks up all the characteristics (signal and noise) that only apply to that set of data. This is called "overfitting." Users won't be able to use a model that's been overfitted for a single dataset when users run it on other datasets.

IV. Building the model: Sometimes, the information or the business goals make sense for a certain algorithm or model. Other times, the simplest way to do something isn't so clear-cut. Users should choose the best model based on what other people have said. Sometimes, users prefer to run a group of models at the same time and choose the best one by comparing their outputs.

V. That process may need to be coordinated with other departments. Build a model that can be used. Even make sure you know how to show ones results to the people who work for the company in a clear and convincing way so that they will adopt ones model. Most models fall apart after a certain amount of time. Keep your model up to date by adding new data to it.

Let’s talk about the questions that we can answer,

RQ1) Can every data be predicted as to what will the outcome be, in the future, ie is it necessary for all kinds of data to be predictable?

Nowadays sophisticated tools have been developed to foresee the future using the data available at hand. Such data tend to have a pattern or a regularity which depends on the variables, making it possible to predict them using mathematical computations. But a very irregular data will lead to probable predictions whose accuracy will be heavily affected. As there is no pattern or dependability, it makes it hard to guess how the data will deviate from its current point. But one can make a model whose accuracy can be improved step by step bring it closer to the predicted result.

RQ2) How should one go forth with establishing a prediction model?

The methodology section answers this question.

3.1) Building a linear regression model in python

The goal of this section is to explain how we can build a model in python using libraries so that we can predict results. In this case we will be using IPL matches datasets to build a model which will be able to identify the behaviour of players performance and give us an approximate score in the first five overs, given that we know the starting line-up and the bowlers playing in the first five overs.

The linear regression equation can be given as :

$Y = m_1x_1 + m_2x_2 + m_3x_3 + \ldots\ldots\ldots\ldots + m_nx_n$

Where $x_{1,\ x2}$,……,x_n are the variables and y is the result which depends on the variables.

In our case we can say Y is the runs in the first five overs, x1, is the average batsmen power (will be discussed ahead), x2 is the average bowlers weakness, x3 is the stadiums effect

So our equation will be :

$Y = m_1x_1 + m_2x_2 + m_3x_3$

Now we have to compute all the variables and their coefficients using python.

Lets discuss, X1: Average batsmen power

We are given a dataset like this:

match_id	season	start_date	venue	innings	ball	batting_team	bowling_team	striker	
335982	2007/08	########	M Chinnas	1	0.1	Kolkata Knight Riders	Royal Challengers Bangal	SC Ganguly	BI
335982	2007/08	########	M Chinnas	1	0.2	Kolkata Knight Riders	Royal Challengers Bangal	BB McCullum	SC
335982	2007/08	########	M Chinnas	1	0.3	Kolkata Knight Riders	Royal Challengers Bangal	BB McCullum	SC
335982	2007/08	########	M Chinnas	1	0.4	Kolkata Knight Riders	Royal Challengers Bangal	BB McCullum	SC
335982	2007/08	########	M Chinnas	1	0.5	Kolkata Knight Riders	Royal Challengers Bangal	BB McCullum	SC
335982	2007/08	########	M Chinnas	1	0.6	Kolkata Knight Riders	Royal Challengers Bangal	BB McCullum	SC
335982	2007/08	########	M Chinnas	1	0.7	Kolkata Knight Riders	Royal Challengers Bangal	BB McCullum	SC
335982	2007/08	########	M Chinnas	1	1.1	Kolkata Knight Riders	Royal Challengers Bangal	BB McCullum	SC
335982	2007/08	########	M Chinnas	1	1.2	Kolkata Knight Riders	Royal Challengers Bangal	BB McCullum	SC
335982	2007/08	########	M Chinnas	1	1.3	Kolkata Knight Riders	Royal Challengers Bangal	BB McCullum	SC
335982	2007/08	########	M Chinnas	1	1.4	Kolkata Knight Riders	Royal Challengers Bangal	BB McCullum	SC
335982	2007/08	########	M Chinnas	1	1.5	Kolkata Knight Riders	Royal Challengers Bangal	BB McCullum	SC
335982	2007/08	########	M Chinnas	1	1.6	Kolkata Knight Riders	Royal Challengers Bangal	BB McCullum	SC
335982	2007/08	########	M Chinnas	1	2.1	Kolkata Knight Riders	Royal Challengers Bangal	SC Ganguly	BI
335982	2007/08	########	M Chinnas	1	2.2	Kolkata Knight Riders	Royal Challengers Bangal	SC Ganguly	BI
335982	2007/08	########	M Chinnas	1	2.3	Kolkata Knight Riders	Royal Challengers Bangal	SC Ganguly	BI
335982	2007/08	########	M Chinnas	1	2.4	Kolkata Knight Riders	Royal Challengers Bangal	BB McCullum	SC
335982	2007/08	########	M Chinnas	1	2.5	Kolkata Knight Riders	Royal Challengers Bangal	BB McCullum	SC
335982	2007/08	########	M Chinnas	1	2.6	Kolkata Knight Riders	Royal Challengers Bangal	SC Ganguly	BI
335982	2007/08	########	M Chinnas	1	3.1	Kolkata Knight Riders	Royal Challengers Bangal	BB McCullum	SC
335982	2007/08	########	M Chinnas	1	3.2	Kolkata Knight Riders	Royal Challengers Bangal	BB McCullum	SC
335982	2007/08	########	M Chinnas	1	3.3	Kolkata Knight Riders	Royal Challengers Bangal	BB McCullum	SC
335982	2007/08	########	M Chinnas	1	3.4	Kolkata Knight Riders	Royal Challengers Bangal	SC Ganguly	BI
335982	2007/08	########	M Chinnas	1	3.5	Kolkata Knight Riders	Royal Challengers Bangal	SC Ganguly	BI
335982	2007/08	########	M Chinnas	1	3.6	Kolkata Knight Riders	Royal Challengers Bangal	SC Ganguly	BI
335982	2007/08	########	M Chinnas	1	3.7	Kolkata Knight Riders	Royal Challengers Bangal	BB McCullum	SC
335982	2007/08	########	M Chinnas	1	4.1	Kolkata Knight Riders	Royal Challengers Bangal	SC Ganguly	BI
335982	2007/08	########	M Chinnas	1	4.2	Kolkata Knight Riders	Royal Challengers Bangal	SC Ganguly	BI
335982	2007/08	########	M Chinnas	1	4.3	Kolkata Knight Riders	Royal Challengers Bangal	BB McCullum	SC
335982	2007/08	########	M Chinnas	1	4.4	Kolkata Knight Riders	Royal Challengers Bangal	BB McCullum	SC

Where each ball has the following details: striker, bowler, runs hit on that ball.

Now since we need to represent entire batting team lineup with one variable, we will have to compute that variable. Let's say each batsman hits a certain amount of runs per ball, that can be given using. Σruns/Σballs. Now this will represent a specific batsman strength. Higher the runs per ball, better the batsmen. Since starting batting team lineup will not be one but more than one player, we will have to take the average of every batsmens strength and hence the variable average batsmen power (X1)

$X1 = \Sigma_{players}(\Sigma runs\_hit/\Sigma balls)$

Similarly, we will compute average bowlers' weakness:

$X2 = \Sigma_{players}(\Sigma runs\_given/\Sigma balls)$

We have an extra detail which indicates which stadium we play the match in, Of course the stadium will affect the runs scored so we can compute the stadiums strength with similar logic. Using pandas, we can represent each player with a number (strength or weakness depending if they are bowler or batsmen):

Batsmen data:

striker	runs_off_bat
A Ashish Reddy	1.428571429
A Chandila	0.571428571
A Chopra	0.706666667
A Choudhary	1.25
A Dananjaya	0.8
A Flintoff	1.087719298
A Kumble	0.714285714
A Mishra	0.882926829
A Mithun	1.307692308
A Mukund	0.826086957
A Nehra	0.650793651
A Nortje	1.166666667
A Singh	0.2
A Symonds	1.247119078
A Uniyal	0.571428571
A Zampa	0.625
AA Bilakhia	0.784090909
AA Chavan	1.090909091
AA Jhunjhunwala	0.995412844
AA Noffke	0.75
AB Agarkar	1.11875
AB Barath	0.976744186
AB Dinda	0.52
AB McDonald	1.194174757
AB de Villiers	1.49014853
AC Blizzard	1.318681319

Bowlers data:

bowler	weakness_coeff
A Ashish Reddy	1.481481481
A Chandila	1.047008547
A Choudhary	1.333333333
A Dananjaya	1.88
A Flintoff	1.606060606
A Kumble	1.107833164
A Mishra	1.211934789
A Mithun	1.528846154
A Nehra	1.2852077
A Nel	1.722222222
A Nortje	1.406914894
A Singh	1.314814815
A Symonds	1.285185185
A Uniyal	1.763157895
A Zampa	1.27739726
AA Chavan	1.329411765
AA Jhunjhunwala	1.477272727
AA Kazi	1.615384615
AA Noffke	1.64
AB Agarkar	1.431707317
AB Dinda	1.323473883
AB McDonald	1.39893617
AC Gilchrist	0
AC Thomas	1.272171254
AC Voges	1.357142857

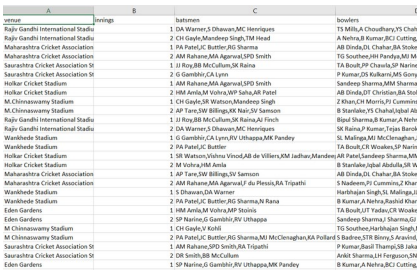
Stadiums Data:

venue	runs_off_bat
Arun Jaitley Stadium	1.296483909
Barabati Stadium	1.294985251
Brabourne Stadium	1.322545053
Buffalo Park	1.037762238
De Beers Diamond Oval	1.181818182
Dr DY Patil Sports Academy	1.124467819
Dr. Y.S. Rajasekhara Reddy ACA-VDCA Cricket Stadi	1.180441225
Dubai International Cricket Stadium	1.233292079
Eden Gardens	1.24766511
Feroz Shah Kotla	1.254408602
Green Park	1.342019544
Himachal Pradesh Cricket Association Stadium	1.25984252
Holkar Cricket Stadium	1.391857506
JSCA International Stadium Complex	1.170556553
Kingsmead	1.118858084
M Chinnaswamy Stadium	1.283893473
M.Chinnaswamy Stadium	1.401259302
MA Chidambaram Stadium	1.156338028
MA Chidambaram Stadium, Chepauk	1.230612245
MA Chidambaram Stadium, Chepauk, Chennai	1.31124498
Maharashtra Cricket Association Stadium	1.278931751
Nehru Stadium	1.113419913
New Wanderers Stadium	1.125773196
Newlands	1.094217024
OUTsurance Oval	0.976
Punjab Cricket Association IS Bindra Stadium	1.372581309

```python
import pandas as pd
import numpy as np
from glob import glob

filenames = glob('./dataset/*.csv')

def append_training_data(filepath):
    df = pd.read_csv(filepath)

    #selecting only 6 overs

    df = df[df['ball']<5.6]

    #removing columns which are not required

    unwanted = ['match_id','season','start_date','batting_team','bowling_team','non_striker','wides','noballs','byes','legbyes','penalty','wicket_type','player_dismi

    df = df.drop(unwanted,axis = 'columns')

    #splitting tables by innings

    innings1 = df[df['innings'] == 1]

    inningsother = df[df['innings'] == 2]

    batsmen1 = innings1.striker.unique().tolist()
    bowlers1 = innings1.bowler.unique().tolist()

    batsmen2 = inningsother.striker.unique().tolist()
    bowlers2 = inningsother.bowler.unique().tolist()

    #making a dataframe for each inning to train the model later
    data1 = [[innings1.venue.iloc[0],1,",".join(batsmen1),",".join(bowlers1),innings1.runs_off_bat.sum() + innings1.extras.sum()]]
    df1 = pd.DataFrame(data1,columns = ['venue','innings','batsmen','bowlers','total'])

    data2 = [[inningsother.venue.iloc[0],2,",".join(batsmen2),",".join(bowlers2),inningsother.runs_off_bat.sum() + inningsother.extras.sum()]]
    df2 = pd.DataFrame(data2,columns = ['venue','innings','batsmen','bowlers','total'])

    #appending to 'training_data.csv'

    df1.to_csv('training_data.csv',mode = 'a',header = False)
    df2.to_csv('training_data.csv',mode = 'a',header = False)
```

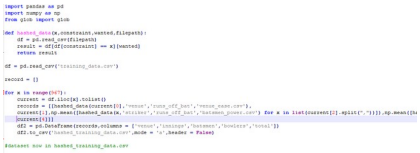
This code is used to extract important information from the given dataset and create a separate excel sheet containing training data which will be used to train our model to find out the coefficients.

The file generated:



venue	innings	batsmen	bowlers
Rajiv Gandhi International Stadiu	1	DA Warner,S Dhawan,MC Henriques	TS Mills,A Choudhary,YS Chah
Rajiv Gandhi International Stadiu	2	CH Gayle,Mandeep Singh,TM Head	A Nehra,B Kumar,BCJ Cutting,
Maharashtra Cricket Association	1	PA Patel,JC Buttler,RG Sharma	AB Dinda,DL Chahar,BA Stoke
Maharashtra Cricket Association	2	AM Rahane,MA Agarwal,SPD Smith	TG Southee,HH Pandya,MJ M
Saurashtra Cricket Association St	1	JJ Roy,BB McCullum,SK Raina	TA Boult,PP Chawla,SP Narine
Saurashtra Cricket Association St	2	G Gambhir,CA Lynn	P Kumar,DS Kulkarni,MS Gony
Holkar Cricket Stadium	1	AM Rahane,MA Agarwal,SPD Smith	Sandeep Sharma,MM Sharma
Holkar Cricket Stadium	2	HM Amla,M Vohra,WP Saha,AR Patel	AB Dinda,DT Christian,BA Stol
M.Chinnaswamy Stadium	1	CH Gayle,SR Watson,Mandeep Singh	Z Khan,CH Morris,PJ Cummin
M.Chinnaswamy Stadium	2	AP Tare,SW Billings,KK Nair,SV Samson	B Stanlake,YS Chahal,Iqbal Ab
Rajiv Gandhi International Stadiu	1	JJ Roy,BB McCullum,SK Raina,AJ Finch	Bipul Sharma,B Kumar,A Nehr
Rajiv Gandhi International Stadiu	2	DA Warner,S Dhawan,MC Henriques	SK Raina,P Kumar,Tejas Baro
Wankhede Stadium	1	G Gambhir,CA Lynn,RV Uthappa,MK Pandey	SL Malinga,MJ McClenaghan,
Wankhede Stadium	2	PA Patel,JC Buttler	TA Boult,CR Woakes,SP Narin
Holkar Cricket Stadium	1	SR Watson,Vishnu Vinod,AB de Villiers,KM Jadhav,Mandeep	AR Patel,Sandeep Sharma,MN
Holkar Cricket Stadium	2	M Vohra,HM Amla	B Stanlake,Iqbal Abdulla,SR W
Maharashtra Cricket Association	1	AP Tare,SW Billings,SV Samson	AB Dinda,DL Chahar,BA Stoke
Maharashtra Cricket Association	2	AM Rahane,MA Agarwal,F du Plessis,RA Tripathi	S Nadeem,PJ Cummins,Z Khar
Wankhede Stadium	1	S Dhawan,DA Warner	Harbhajan Singh,SL Malinga,JJ
Wankhede Stadium	2	PA Patel,JC Buttler,RG Sharma,N Rana	B Kumar,A Nehra,Rashid Khar
Eden Gardens	1	HM Amla,M Vohra,MP Stoinis	TA Boult,IJT Yadav,CR Woake
Eden Gardens	2	SP Narine,G Gambhir,RV Uthappa	Sandeep Sharma,J Sharma,GJ
M Chinnaswamy Stadium	1	CH Gayle,V Kohli	TG Southee,Harbhajan Singh,J
M Chinnaswamy Stadium	2	PA Patel,JC Buttler,RG Sharma,MJ McClenaghan,KA Pollard	S Badree,STR Binny,S Aravind,
Saurashtra Cricket Association St	1	AM Rahane,SPD Smith,RA Tripathi	P Kumar,Basil Thampi,SB Jaka
Saurashtra Cricket Association St	2	DR Smith,BB McCullum	Ankit Sharma,LH Ferguson,SN
Eden Gardens	1	SP Narine,G Gambhir,RV Uthappa,MK Pandey	B Kumar,A Nehra,BCJ Cutting,

This sheet contains teams lineups and the total score. We can use this data to train our model using existing linear regression model tools.



```
import pandas as pd
import numpy as np
from glob import glob

def hashed_data(m,constraint,wanted,filepath):
    df = pd.read_csv(filepath)
    result = df[df[constraint] == m][wanted]
    return result

df = pd.read_csv('training_data.csv')

record = []

for m in range(747):
    current = df.iloc[m].tolist()
    records = [[hashed_data(current[2],'venue','runs_off_bat','venue_name.csv'),
    current[1],np.mean([hashed_data(m,'striker','runs_off_bat','batsmen_scores.csv') for m in list(current[3].split(","))]),np.mean([ha
    current[4]]]
    df2 = pd.DataFrame(records,columns = ["venue","innings","batsmen","bowlers","total"])
    df2.to_csv('hashed_training_data.csv',mode = 'a',header = False)

#dataset now in hashed_training_data.csv
```

This code is used to transform the team lineups into a single number which represent the average weight of the players strength or weakness.

venue	innings	batsmen	bowlers
1.111455108	1	1.190469872	1.25836835
1.111455108	2	0.95069409	1.22986284
1.200131666	1	1.254690797	1.38084487
1.200131666	2	1.163145033	1.30606368
1.427012278	1	1.223374356	1.27668019
1.427012278	2	1.290104415	1.45032286
1.312788906	1	1.163145033	1.24536451
1.312788906	2	1.213955039	1.38158276
1.295774648	1	1.172992079	1.23188689
1.295774648	2	1.165483295	1.2409548
1.111455108	1	1.1948641	1.10046309
1.111455108	2	1.190469872	1.23946729
1.127447289	1	1.226178144	1.13617650
1.127447289	2	1.334016393	1.27225967
1.312788906	1	0.943189241	1.27619845
1.312788906	2	1.251203251	1.18714918
1.200131666	1	1.151678877	1.29210689
1.200131666	2	1.231466299	1.23188689
1.127447289	1	1.252371474	1.09192331
1.127447289	2	1.224968715	1.09213955
1.155061019	1	1.186408228	1.27565841
1.155061019	2	1.354293889	1.33123797
1.112865372	1	1.20837507	1.24898270
1.112865372	2	0.979608773	1.24741450
1.427012278	1	1.23399238	1.3664391
1.427012278	2	1.209429503	1.41936618

Now the final step is to train the model:

```
import pickle
import pandas as pd
import numpy as np
import math as ma

df = pd.read_csv('inputFile.csv')
with open('model_pickle','rb') as f:
    mp = pickle.load(f)

def hashed_data(m,constraint,wanted,filepath):
    df = pd.read_csv(filepath)
    result = df[df[constraint] == m][wanted]
    return result

def predict(predictArray):
    return mp.predict(predictArray)

def predictRuns(testInputs):
    prediction = 0

    current = df.iloc[0].tolist()

    batsmen = current[3].split(',')
    bowlers = current[4].split(',')

    predictArray = [[int(hashed_data(current[2],'venue','runs_off_bat','./used_for_hashing/venue_name.csv')),current[1],np.mean([hashed_data(m,'striker'
    hashed_data(m,'bowler','wickets_confF','./used_for_hashing/bowlers_weakness.csv') for m in bowlers)]]]
    prediction = predict(predictArray)

    return ma.floor(prediction)
```

We use a module called pickle which trains the model, given the data. We have to specify the variables and the output variable. After training, it saves the model in a pickle file which we can use to load the data and predict using this pretrained pickle file. Let's see it in action.

This is the input file:

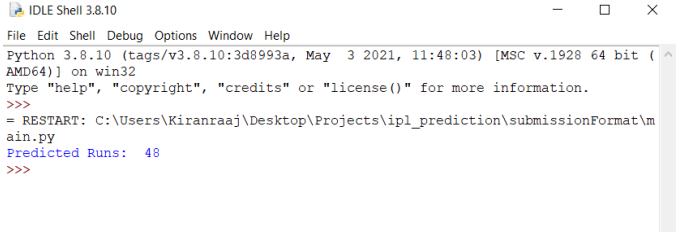
inputFile.csv:

venue	innings	batting_team	bowling_team	batsmen	bowlers
Arun Jaitle	1	Rajasthan Royals	Mumbai Indians	JC Buttler,YBK Jaiswal	TA Boult,J Yadav,JJ Bumrah,NM Coulter-Nile

After running a py file which loads up the pickle file and puts in this input data, we will be shown the results

```
from predictor import predictRuns


"""
sys.argv[1] is the input test file name given as command line arguments

"""
runs = predictRuns('inputFile.csv')
print("Predicted Runs: ", runs)
```

Output:



```
IDLE Shell 3.8.10
File  Edit  Shell  Debug  Options  Window  Help
Python 3.8.10 (tags/v3.8.10:3d8993a, May  3 2021, 11:48:03) [MSC v.1928 64 bit (
AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
= RESTART: C:\Users\Kiranraaj\Desktop\Projects\ipl_prediction\submissionFormat\m
ain.py
Predicted Runs:  48
>>>
```

Hence, readers are suggested to read articles [6-13] to know about importance of Machine learning, deep learning techniques for analysis big a data (with explaining few uses cases of real-world problem), which will help researchers to find out an efficient solution for their research work.

4. Conclusion

Big data is a potential research area receiving considerable attention from academia and IT communities. In the digital world, the amounts of data generated and stored have expanded within a short period of time. Consequently, this fast-growing rate of data has created many challenges[R5]. The way we live and do business will be changed due to data analysis. Although organizations are taking steps to show data into insights, our global survey showed that organizations are still battling data quality and therefore the problem to seek out the proper resources to show these insights into true value and become more data-driven. What's the current condition and possibilities in the future from the information era. Stages of data analytics are best known as:: from descriptive, to diagnostic, to discovery , to predictive and, finally, to data analytics (what action is that the best to take).. If it seems within the future that a decision-making process supported data analytics will produce better results, the step to “automated” decision-making are going to be small like AI Examples are the autopilot update within the Tesla model S cars, which has been driving around for almost a million miles and that too without ever getting a ticket. The question is will this transform the way we live? The answer is yes! Elon musk the founder of Tesla cars says that at some future point it would not make sense to drive a car, even though this sounds remotely close to insanse but mind it we even thought that we could survive without cell phones! Developments tend to travel an extended way towards a situation where, for instance, you'll reduce your automobile insurance premium once you share all sensor data of your automobile with the insurance firm. This is not possible now but hopefully in the mere future.

References

1. Runkler, T. A. (2020). *Data analytics*. Springer Fachmedien Wiesbaden.
2. Wright, Raymond E. "Logistic regression." (1995).
3. Marquardt, Donald W., and Ronald D. Snee. "Ridge regression in practice." *The American Statistician* 29.1 (1975): 3-20.
4. Albawi, Saad, Tareq Abed Mohammed, and Saad Al-Zawi. "Understanding of a convolutional neural network." *2017 International Conference on Engineering and Technology (ICET)*. Ieee, 2017.
5. Yaqoob, I., Hashem, I. A. T., Gani, A., Mokhtar, S., Ahmed, E., Anuar, N. B., & Vasilakos, A. V. (2016). Big data: From beginning to future. *International Journal of Information Management*, *36*(6), 1231-1247.
6. Kumari S., Vani V., Malik S., Tyagi A.K., Reddy S. (2021) Analysis of Text Mining Tools in Disease Prediction. In: Abraham A., Hanne T., Castillo O., Gandhi N., Nogueira Rios T., Hong TP. (eds) Hybrid Intelligent Systems. HIS 2020. Advances in Intelligent Systems and Computing, vol 1375. Springer, Cham. https://doi.org/10.1007/978-3-030-73050-5_55

7. B. Gudeti, S. Mishra, S. Malik, T. F. Fernandez, A. K. Tyagi and S. Kumari, "A Novel Approach to Predict Chronic Kidney Disease using Machine Learning Algorithms," 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, 2020, pp. 1630-1635, doi: 10.1109/ICECA49313.2020.9297392.
8. Amit Kumar Tyagi, Dr. Meenu Gupta, Aswathy SU, Chetanya Ved, “Healthcare Solutions for Smart Era: An Useful Explanation from User’s Perspective”, in the Book "Recent Trends in Blockchain for Information Systems Security and Privacy", CRC Press, 2021.
9. L. Kanuru, A. K. Tyagi, A. S. U, T. F. Fernandez, N. Sreenath and S. Mishra, "Prediction of Pesticides and Fertilizers using Machine Learning and Internet of Things," 2021 International Conference on Computer Communication and Informatics (ICCCI), 2021, pp. 1-6, doi: 10.1109/ICCCI50826.2021.9402536.
10. Amit Kumar Tyagi, Poonam Chahal, “Artificial Intelligence and Machine Learning Algorithms”, Book: Challenges and Applications for Implementing Machine Learning in Computer Vision, IGI Global, 2020.DOI: 10.4018/978-1-7998-0182-5.ch008
11. Amit Kumar Tyagi, G. Rekha, “Challenges of Applying Deep Learning in Real-World Applications”, Book: Challenges and Applications for Implementing Machine Learning in Computer Vision, IGI Global 2020, p. 92-118. DOI: 10.4018/978-1-7998-0182-5.ch004
12. Akshara Pramod, Harsh Sankar Naicker, Amit Kumar Tyagi, “Machine Learning and Deep Learning: Open Issues and Future Research Directions for Next Ten Years”, Book: Computational Analysis and Understanding of Deep Learning for Medical Care: Principles, Methods, and Applications, 2020, Wiley Scrivener, 2020.
13. Tyagi, Amit Kumar and G, Rekha, Machine Learning with Big Data (March 20, 2019). Proceedings of International Conference on Sustainable Computing in Science, Technology and Management (SUSCOM), Amity University Rajasthan, Jaipur - India, February 26-28, 2019.