Study of data behaviour and methods for data prediction and analysis

Kiranraaj Sekar (20BCE1418)

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.

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 mans term sounds like its 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 gadgets. 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

And as a result, as data is wealth. The data analysis job sector started rising through the roof. Information Science and Data Analytics are two popular expressions of the year. Today, information is more than oil to the businesses. Information is gathered into crude structure and handled by the necessity of an organization and afterward take this information for the dynamic reason. This cycle, assists the business with filling on the lookout. Yet, who will do this work? Who will handle the information? and so on Everything is finished by a Data Analytics and a Data Scientist. Normal occupation development across all occupations in the U.S. is 3.7%. There's more proof of a developing interest for information examination experts, particularly among chiefs and hierarchical pioneers. 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 behaviour 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 assosications 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 amount 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:

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?

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

Keyword/s: EB – Exabytes.

2) Related work

There are many different types of predictive modeling techniques including ANOVA, linear regression (ordinary least squares), logistic regression, ridge regression, time series, decision trees, neural networks, and many more. Selecting the correct predictive modeling technique at the start of your project can save a lot of time. Choosing the incorrect modeling technique can result in inaccurate predictions and residual plots that experience non-constant variance and/or mean.

2.1) Regression Analysis

[R1]Regression estimates functional dependencies between features. Linear regression models can be efficiently computed from covariances but are restricted to linear dependencies.Regression analysis is employed to predict endless target variable from one or multiple independent variables. Regression analysis is used with natural variables, rather than variables that change through experimentation. As stated above, there are many various sorts of regression, so once we've decided multivariate analysis should be used, how can we choose which regression technique should be applied?

2.1.1) Anova

Anova comes into picture when the target final variables happen to be continuos and the given input variables are categorical and discontinuous. Null hypotheses considers all groups as equal and assumes that the data should be somewhat similarly distributed hence meaning that every other sample case will approximately have 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 variale is linearly dependent on the input variables. The input variables can be continuous or discontinuous meaning categorical.

2.1.3) Logistic Regression

[R2] Logistic regression 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 a method to predict future responses supported by response history. The information for a statistic must be a set of observations about the values that a variable takes at different times. the information is bivariate and, therefore, the experimental variable is time. The rows must be stationary; H. are normally distributed: the mean value and variance of the series are constant over long periods of time. Furthermore, the residuals must be normally distributed even over a longer period of time, even if they are not correlated. The series must not contain outliers. In fact, if there are random shocks, they should be randomized with a mean of 0 and a sustained variance.

2.1.5) Ridge Regression

[R3]. The use of biased estimation in data analysis and model building is discussed. A review of the theory of ridge regression and its relation to generalized inverse regression is presented along with the results of a simulation experiment and three examples of the use of ridge regression in practice. Comments on variable selection procedures, model validation, and ridge and generalized inverse regression computation procedures are included. The examples studied here show that when the predictor variables are highly correlated, ridge regression produces coefficients which predict and extrapolate better than least squares and is a safe procedure for selecting 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. 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

[R4] The term Deep Learning or Deep Neural Network refers to Artificial Neural Networks (ANN) with multi layers. Over the last few decades, it has been considered to be one of the most powerful tools, and has become very popular in the literature as it is able to handle a huge amount of data. The interest in having deeper hidden layers has recently begun to surpass classical methods performance in different fields; especially in pattern recognition. One of the most popular deep neural networks is the Convolutional Neural Network (CNN). It take this name from mathematical linear operation between matrixes called convolution. CNN have multiple layers; including convolutional layer, non-linearity layer, pooling layer and fully-connected layer. The convolutional and fully-connected layers have parameters but pooling and non-linearity layers don't have parameters. The CNN has an excellent performance in machine learning problems. Specially the applications that deal with image data, such as largest image classification data set

3)Methodology

Following are the points jotted down as a guideline for building a predictive model

1) Preparing business objectives

The info or the business objectives lend themselves to a chosen algorithm or model. Other times the simplest approach isn't so clear-cut. As you explore the info , run as many algorithms as you can; compare their outputs. Base your choice of the ultimate model on the general results. Sometimes you're more happy running an ensemble of models simultaneously on the info and selecting a final model by comparing their outputs.

2) Preparing data

You'll use historical data to coach your model. the info is typically scattered across multiple sources and should require cleansing and preparation. Data may contain duplicate records and outliers; counting on the analysis and therefore the business objective, you opt whether to stay or remove them. Also, the info could have missing values, may have to undergo some transformation, and should be wont to generate derived attributes that have more predictive power for your objective. Overall, the standard of the info indicates the standard of the model.

3)Sampling your data

You'll got to split your data into two sets: training and test datasets. You build the model using the training dataset. you employ the test data set to verify the accuracy of the model's output. Doing so is completely crucial. Otherwise you run the danger of overfitting your model — training the model with a limited dataset, to the purpose that it picks all the characteristics (both the signal and therefore the noise) that are only true for that specific dataset. An model that's overfitted for a selected data set will perform miserably once you run it on other datasets. A test dataset ensures a legitimate thanks to accurately measure your model's performance.

4) Building the model

Sometimes the info or the business objectives lend themselves to a selected algorithm or model. Other times the simplest approach isn't so clear-cut. As you explore the info , run as many algorithms as you can; compare their outputs. Base your choice of the ultimate model on the general results. Sometimes you're more happy running an ensemble of models simultaneously on the info and selecting a final model by comparing their outputs.

5) Deploying the model

After building the model, you've got to deploy it so as to reap its benefits. That process may require co-ordination with other departments. Aim at building a deployable model. even be sure you recognize the way to present your results to the business stakeholders in a clear and convincing way in order that they adopt your model. After the model is deployed, you'll got to monitor its performance and continue improving it. Most models decay after a particular period of your time . Keep your model up so far by refreshing it with newly available data.

Lets 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 ,x_2 ,………….,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

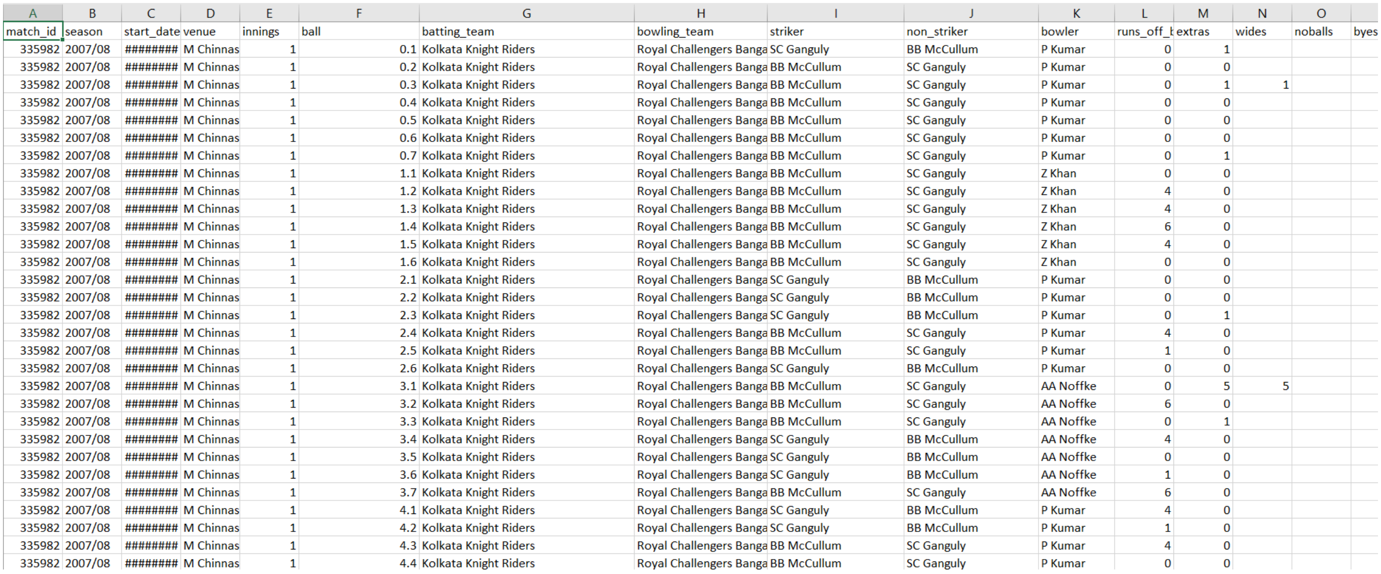
So our equation will be :

$$Y = m_1x_1 + m_2x_2 + m_3x_3$$

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	non_striker	bowler	runs_off_b	extras	wides	noballs	byes
335982	2007/08	########	M Chinnas	1	0.1	Kolkata Knight Riders	Royal Challengers Banga	SC Ganguly	BB McCullum	P Kumar	0	1			
335982	2007/08	########	M Chinnas	1	0.2	Kolkata Knight Riders	Royal Challengers Banga	BB McCullum	SC Ganguly	P Kumar	0	0			
335982	2007/08	########	M Chinnas	1	0.3	Kolkata Knight Riders	Royal Challengers Banga	BB McCullum	SC Ganguly	P Kumar	0	1	1		
335982	2007/08	########	M Chinnas	1	0.4	Kolkata Knight Riders	Royal Challengers Banga	BB McCullum	SC Ganguly	P Kumar	0	0			
335982	2007/08	########	M Chinnas	1	0.5	Kolkata Knight Riders	Royal Challengers Banga	BB McCullum	SC Ganguly	P Kumar	0	0			
335982	2007/08	########	M Chinnas	1	0.6	Kolkata Knight Riders	Royal Challengers Banga	BB McCullum	SC Ganguly	P Kumar	0	0			
335982	2007/08	########	M Chinnas	1	0.7	Kolkata Knight Riders	Royal Challengers Banga	BB McCullum	SC Ganguly	P Kumar	0	1			
335982	2007/08	########	M Chinnas	1	1.1	Kolkata Knight Riders	Royal Challengers Banga	BB McCullum	SC Ganguly	Z Khan	0	0			
335982	2007/08	########	M Chinnas	1	1.2	Kolkata Knight Riders	Royal Challengers Banga	BB McCullum	SC Ganguly	Z Khan	4	0			
335982	2007/08	########	M Chinnas	1	1.3	Kolkata Knight Riders	Royal Challengers Banga	BB McCullum	SC Ganguly	Z Khan	4	0			
335982	2007/08	########	M Chinnas	1	1.4	Kolkata Knight Riders	Royal Challengers Banga	BB McCullum	SC Ganguly	Z Khan	6	0			
335982	2007/08	########	M Chinnas	1	1.5	Kolkata Knight Riders	Royal Challengers Banga	BB McCullum	SC Ganguly	Z Khan	4	0			
335982	2007/08	########	M Chinnas	1	1.6	Kolkata Knight Riders	Royal Challengers Banga	BB McCullum	SC Ganguly	Z Khan	0	0			
335982	2007/08	########	M Chinnas	1	2.1	Kolkata Knight Riders	Royal Challengers Banga	SC Ganguly	BB McCullum	P Kumar	0	0			
335982	2007/08	########	M Chinnas	1	2.2	Kolkata Knight Riders	Royal Challengers Banga	SC Ganguly	BB McCullum	P Kumar	0	0			
335982	2007/08	########	M Chinnas	1	2.3	Kolkata Knight Riders	Royal Challengers Banga	SC Ganguly	BB McCullum	P Kumar	0	1			
335982	2007/08	########	M Chinnas	1	2.4	Kolkata Knight Riders	Royal Challengers Banga	BB McCullum	SC Ganguly	P Kumar	4	0			
335982	2007/08	########	M Chinnas	1	2.5	Kolkata Knight Riders	Royal Challengers Banga	BB McCullum	SC Ganguly	P Kumar	1	0			
335982	2007/08	########	M Chinnas	1	2.6	Kolkata Knight Riders	Royal Challengers Banga	SC Ganguly	BB McCullum	P Kumar	0	0			
335982	2007/08	########	M Chinnas	1	3.1	Kolkata Knight Riders	Royal Challengers Banga	BB McCullum	SC Ganguly	AA Noffke	0	5	5		
335982	2007/08	########	M Chinnas	1	3.2	Kolkata Knight Riders	Royal Challengers Banga	BB McCullum	SC Ganguly	AA Noffke	6	0			
335982	2007/08	########	M Chinnas	1	3.3	Kolkata Knight Riders	Royal Challengers Banga	BB McCullum	SC Ganguly	AA Noffke	0	1			
335982	2007/08	########	M Chinnas	1	3.4	Kolkata Knight Riders	Royal Challengers Banga	SC Ganguly	BB McCullum	AA Noffke	4	0			
335982	2007/08	########	M Chinnas	1	3.5	Kolkata Knight Riders	Royal Challengers Banga	SC Ganguly	BB McCullum	AA Noffke	0	0			
335982	2007/08	########	M Chinnas	1	3.6	Kolkata Knight Riders	Royal Challengers Banga	SC Ganguly	BB McCullum	AA Noffke	1	0			
335982	2007/08	########	M Chinnas	1	3.7	Kolkata Knight Riders	Royal Challengers Banga	BB McCullum	SC Ganguly	AA Noffke	6	0			
335982	2007/08	########	M Chinnas	1	4.1	Kolkata Knight Riders	Royal Challengers Banga	SC Ganguly	BB McCullum	P Kumar	4	0			
335982	2007/08	########	M Chinnas	1	4.2	Kolkata Knight Riders	Royal Challengers Banga	SC Ganguly	BB McCullum	P Kumar	1	0			
335982	2007/08	########	M Chinnas	1	4.3	Kolkata Knight Riders	Royal Challengers Banga	BB McCullum	SC Ganguly	P Kumar	4	0			
335982	2007/08	########	M Chinnas	1	4.4	Kolkata Knight Riders	Royal Challengers Banga	BB McCullum	SC Ganguly	P Kumar	0	0			

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. Lets say each batsman hits a certain amount of runs per ball, That can be given using :

$\Sigma runs/\Sigma balls$. 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)

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

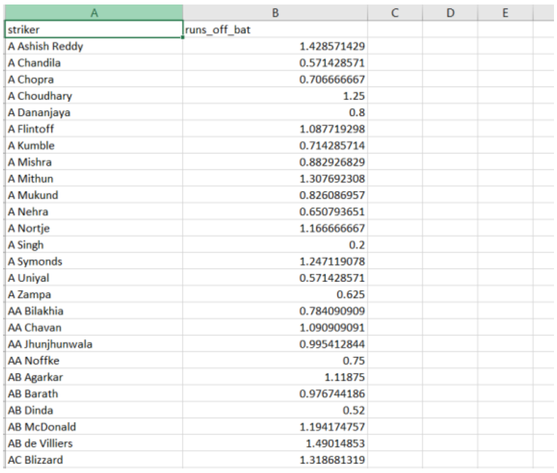
Similarly we will compute average bowlers weakness:

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

We have an extra detail which indicates which stadium we play the match in, Ofcourse the stadium will affect the runs scored so we can compute the stadiums strength with similar logic.

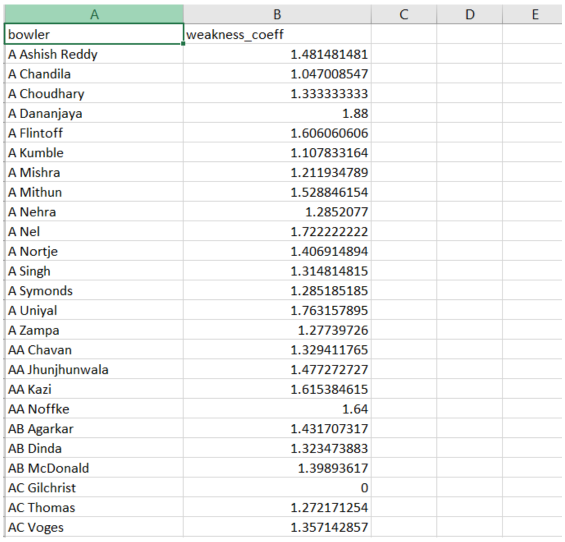
Using pandas we can represent each player with a number (strength or weakness depending if they are bowler or batsmen):

Batsmen data:



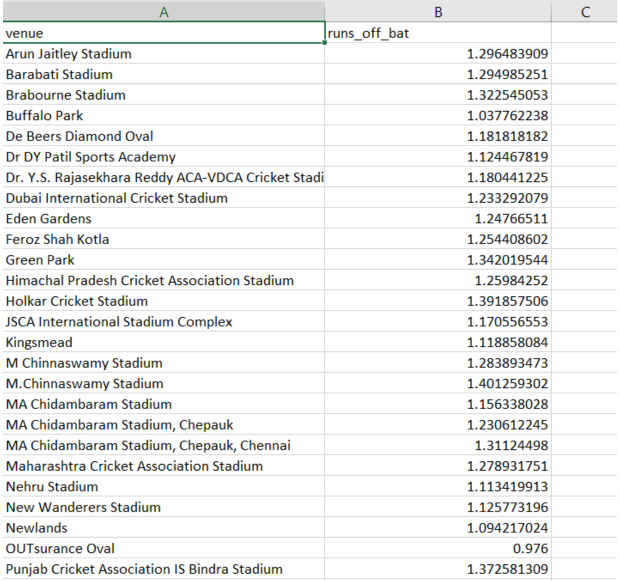
striker	runs_off_bat
A Ashish Reddy	1.428571429
A Chandila	0.571428571
A Chopra	0.706666667
A Choudhary	1.25
A Dananjaya	0.8
A Flintoff	1.087719298
A Kumble	0.714285714
A Mishra	0.882926829
A Mithun	1.307692308
A Mukund	0.826086957
A Nehra	0.650793651
A Nortje	1.166666667
A Singh	0.2
A Symonds	1.247119078
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:

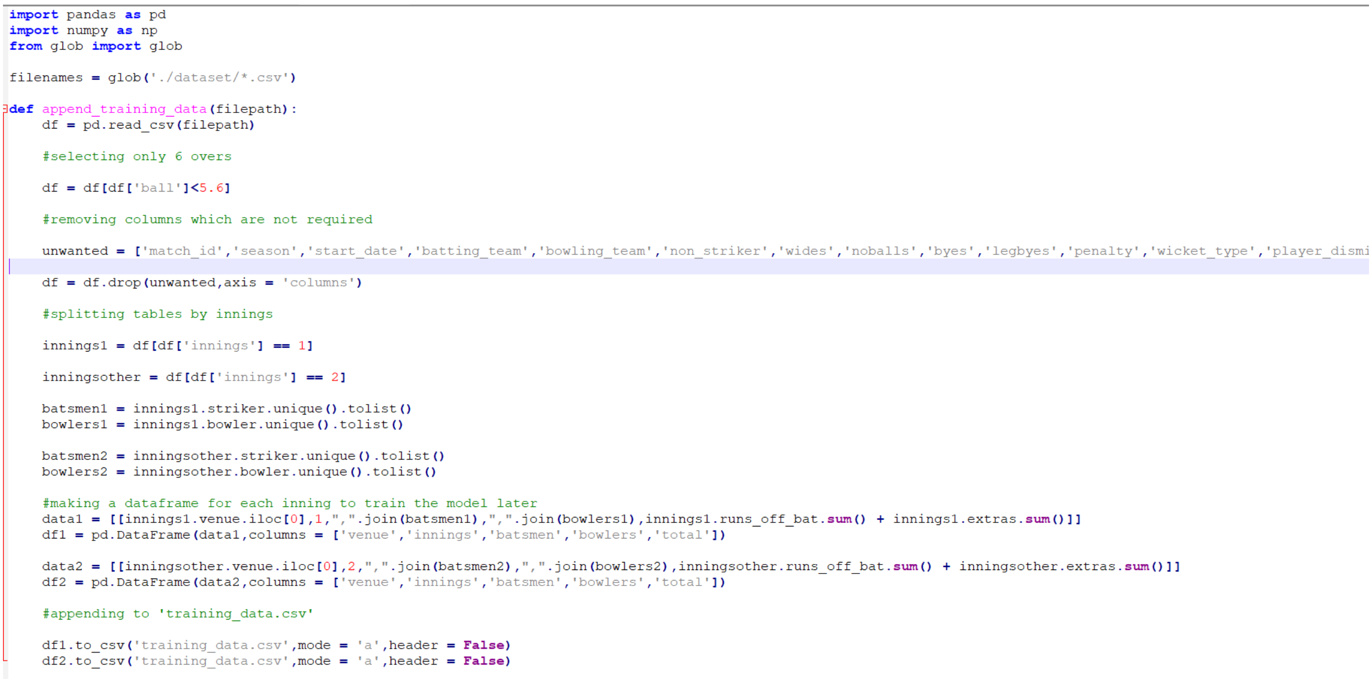


bowler	weakness_coeff
A Ashish Reddy	1.481481481
A Chandila	1.047008547
A Choudhary	1.333333333
A Dananjaya	1.88
A Flintoff	1.606060606
A Kumble	1.107833164
A Mishra	1.211934789
A Mithun	1.528846154
A Nehra	1.2852077
A Nel	1.722222222
A Nortje	1.406914894
A Singh	1.314814815
A Symonds	1.285185185
A Uniyal	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:



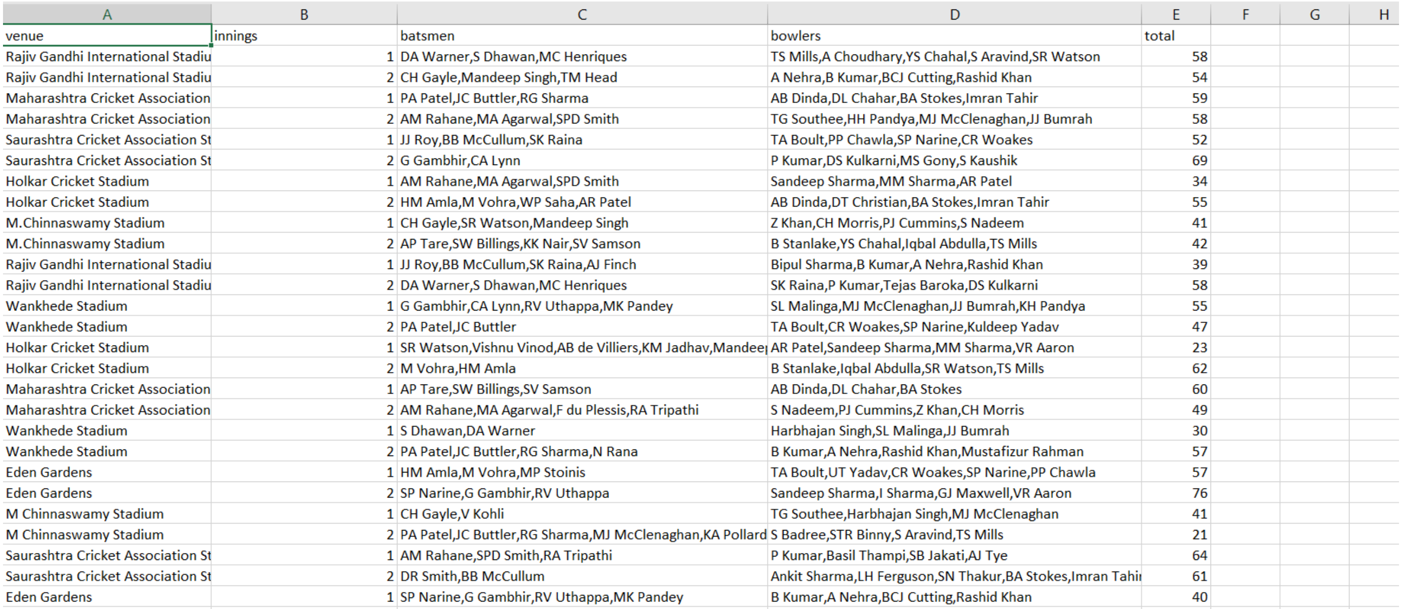
venue	runs_off_bat
Arun Jaitley Stadium	1.296483909
Barabati Stadium	1.294985251
Brabourne Stadium	1.322545053
Buffalo Park	1.037762238
De Beers Diamond Oval	1.181818182
Dr DY Patil Sports Academy	1.124467819
Dr. Y.S. Rajasekhara Reddy ACA-VDCA Cricket Stadi	1.180441225
Dubai International Cricket Stadium	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



```
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)
```

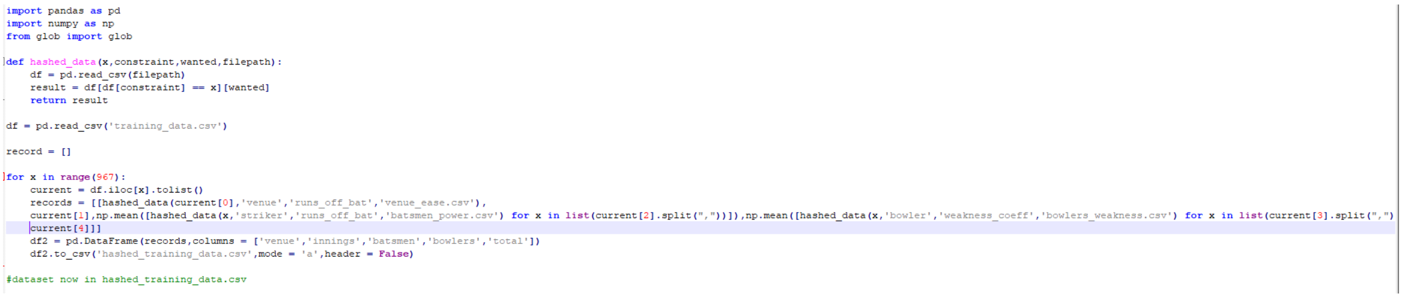
This code is used to extract important information from the given dataset and create a separate excel sheet containing training data which will be used to train our model to find out the coefficients.

The file generated:



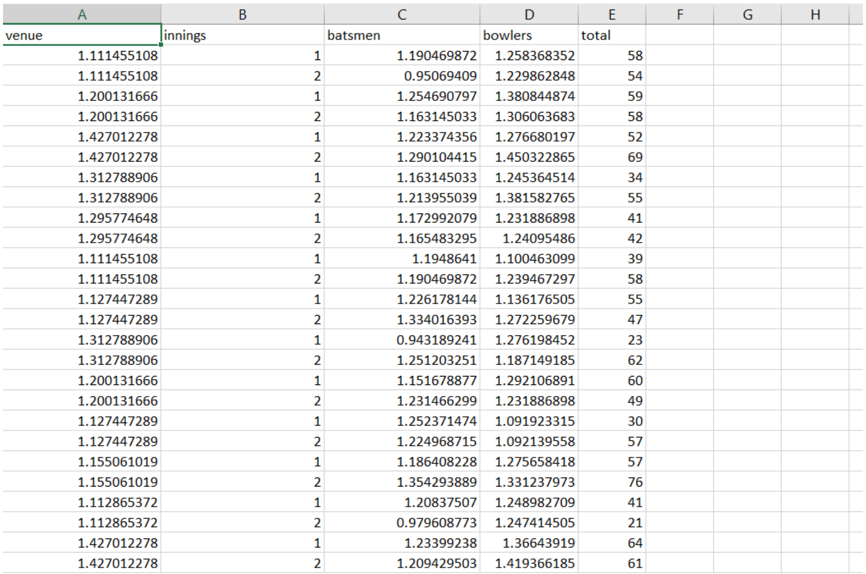
A	B	C	D	E
venue	innings	batsmen	bowlers	total
Rajiv Gandhi International Stadiu	1	DA Warner,S Dhawan,MC Henriques	TS Mills,A Choudhary,YS Chahal,S Aravind,SR Watson	58
Rajiv Gandhi International Stadiu	2	CH Gayle,Mandeep Singh,TM Head	A Nehra,B Kumar,BCJ Cutting,Rashid Khan	54
Maharashtra Cricket Association	1	PA Patel,JC Buttler,RG Sharma	AB Dinda,DL Chahar,BA Stokes,Imran Tahir	59
Maharashtra Cricket Association	2	AM Rahane,MA Agarwal,SPD Smith	TG Southee,HH Pandya,MJ McClenaghan,JJ Bumrah	58
Saurashtra Cricket Association St	1	JJ Roy,BB McCullum,SK Raina	TA Boult,PP Chawla,SP Narine,CR Woakes	52
Saurashtra Cricket Association St	2	G Gambhir,CA Lynn	P Kumar,DS Kulkarni,MS Gony,S Kaushik	69
Holkar Cricket Stadium	1	AM Rahane,MA Agarwal,SPD Smith	Sandeep Sharma,MM Sharma,AR Patel	34
Holkar Cricket Stadium	2	HM Amla,M Vohra,WP Saha,AR Patel	AB Dinda,DT Christian,BA Stokes,Imran Tahir	55
M.Chinnaswamy Stadium	1	CH Gayle,SR Watson,Mandeep Singh	Z Khan,CH Morris,PJ Cummins,S Nadeem	41
M.Chinnaswamy Stadium	2	AP Tare,SW Billings,KK Nair,SV Samson	B Stanlake,YS Chahal,Iqbal Abdulla,TS Mills	42
Rajiv Gandhi International Stadiu	1	JJ Roy,BB McCullum,SK Raina,AJ Finch	Bipul Sharma,B Kumar,A Nehra,Rashid Khan	39
Rajiv Gandhi International Stadiu	2	DA Warner,S Dhawan,MC Henriques	SK Raina,P Kumar,Tejas Baroka,DS Kulkarni	58
Wankhede Stadium	1	G Gambhir,CA Lynn,RV Uthappa,MK Pandey	SL Malinga,MJ McClenaghan,JJ Bumrah,KH Pandya	55
Wankhede Stadium	2	PA Patel,JC Buttler	TA Boult,CR Woakes,SP Narine,Kuldeep Yadav	47
Holkar Cricket Stadium	1	SR Watson,Vishnu Vinod,AB de Villiers,KM Jadhav,Mandee	AR Patel,Sandeep Sharma,MM Sharma,VR Aaron	23
Holkar Cricket Stadium	2	M Vohra,HM Amla	B Stanlake,Iqbal Abdulla,SR Watson,TS Mills	62
Maharashtra Cricket Association	1	AP Tare,SW Billings,SV Samson	AB Dinda,DL Chahar,BA Stokes	60
Maharashtra Cricket Association	2	AM Rahane,MA Agarwal,F du Plessis,RA Tripathi	S Nadeem,PJ Cummins,Z Khan,CH Morris	49
Wankhede Stadium	1	S Dhawan,DA Warner	Harbhajan Singh,SL Malinga,JJ Bumrah	30
Wankhede Stadium	2	PA Patel,JC Buttler,RG Sharma,N Rana	B Kumar,A Nehra,Rashid Khan,Mustafizur Rahman	57
Eden Gardens	1	HM Amla,M Vohra,MP Stoinis	TA Boult,UT Yadav,CR Woakes,SP Narine,PP Chawla	57
Eden Gardens	2	SP Narine,G Gambhir,RV Uthappa	Sandeep Sharma,I Sharma,GJ Maxwell,VR Aaron	76
M Chinnaswamy Stadium	1	CH Gayle,V Kohli	TG Southee,Harbhajan Singh,MJ McClenaghan	41
M Chinnaswamy Stadium	2	PA Patel,JC Buttler,RG Sharma,MJ McClenaghan,KA Pollard	S Badree,STR Binny,S Aravind,TS Mills	21
Saurashtra Cricket Association St	1	AM Rahane,SPD Smith,RA Tripathi	P Kumar,Basil Thampi,SB Jakati,AJ Tye	64
Saurashtra Cricket Association St	2	DR Smith,BB McCullum	Ankit Sharma,LH Ferguson,SN Thakur,BA Stokes,Imran Tahi	61
Eden Gardens	1	SP Narine,G Gambhir,RV Uthappa,MK Pandey	B Kumar,A Nehra,BCJ Cutting,Rashid Khan	40

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



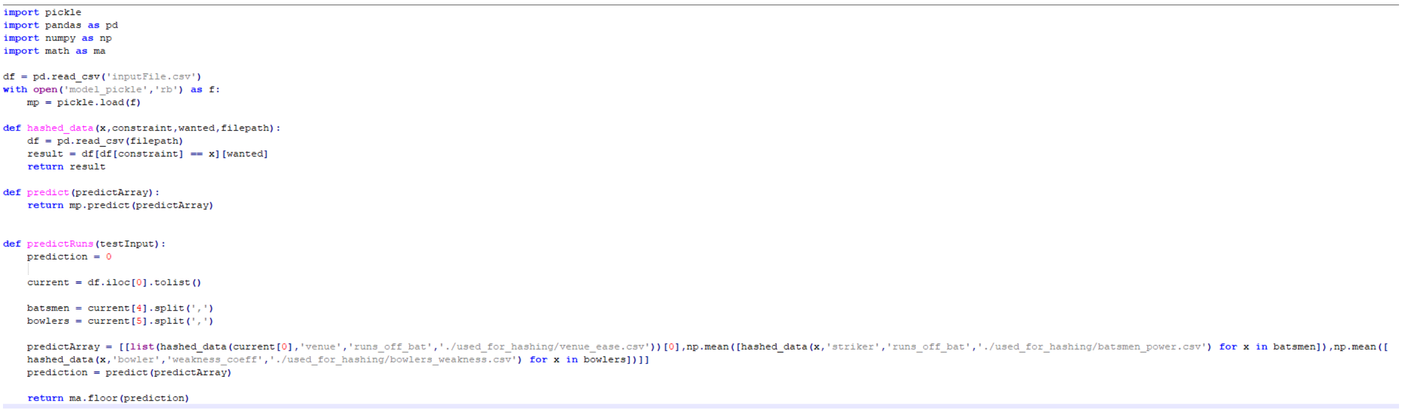
```
import pandas as pd
import numpy as np
from glob import glob

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

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

record = []

for x in range(967):
    current = df.iloc[x].tolist()
    records = [[hashed_data(current[0],'venue','runs_off_bat','venue_ease.csv'),
    current[1],np.mean([hashed_data(x,'striker','runs_off_bat','batsmen_power.csv') for x in list(current[2].split(","))]),np.mean([hashed_data(x,'bowler','weakness_coeff','bowlers_weakness.csv') for x in list(current[3].split(",")
    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.



A	B	C	D	E
venue	innings	batsmen	bowlers	total
1.111455108	1	1.190469872	1.258368352	58
1.111455108	2	0.95069409	1.229862848	54
1.200131666	1	1.254690797	1.380844874	59
1.200131666	2	1.163145033	1.306063683	58
1.427012278	1	1.223374356	1.276680197	52
1.427012278	2	1.290104415	1.450322865	69
1.312788906	1	1.163145033	1.245364514	34
1.312788906	2	1.213955039	1.381582765	55
1.295774648	1	1.172992079	1.231886898	41
1.295774648	2	1.165483295	1.24095486	42
1.111455108	1	1.1948641	1.100463099	39
1.111455108	2	1.190469872	1.239467297	58
1.127447289	1	1.226178144	1.136176505	55
1.127447289	2	1.334016393	1.272259679	47
1.312788906	1	0.943189241	1.276198452	23
1.312788906	2	1.251203251	1.187149185	62
1.200131666	1	1.151678877	1.292106891	60
1.200131666	2	1.231466299	1.231886898	49
1.127447289	1	1.252371474	1.091923315	30
1.127447289	2	1.224968715	1.092139558	57
1.155061019	1	1.186408228	1.275658418	57
1.155061019	2	1.354293889	1.331237973	76
1.112865372	1	1.20837507	1.248982709	41
1.112865372	2	0.979608773	1.247414505	21
1.427012278	1	1.23399238	1.36643919	64
1.427012278	2	1.209429503	1.419366185	61

Now the final step is to train the model:

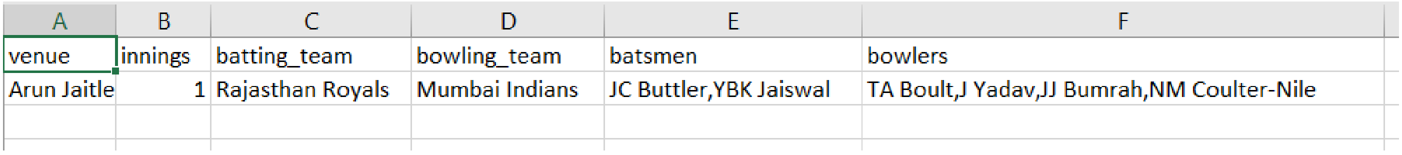


```
import pickle
import pandas as pd
import numpy as np
import math as ma

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

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

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


def predictRuns(testInput):
    prediction = 0

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

    batsmen = current[4].split(',')
    bowlers = current[5].split(',')

    predictArray = [[list(hashed_data(current[0],'venue','runs_off_bat','./used_for_hashing/venue_ease.csv'))[0],np.mean([hashed_data(x,'striker','runs_off_bat','./used_for_hashing/batsmen_power.csv') for x in batsmen]),np.mean([
    hashed_data(x,'bowler','weakness_coeff','./used_for_hashing/bowlers_weakness.csv') for x 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. Lets see it in action.

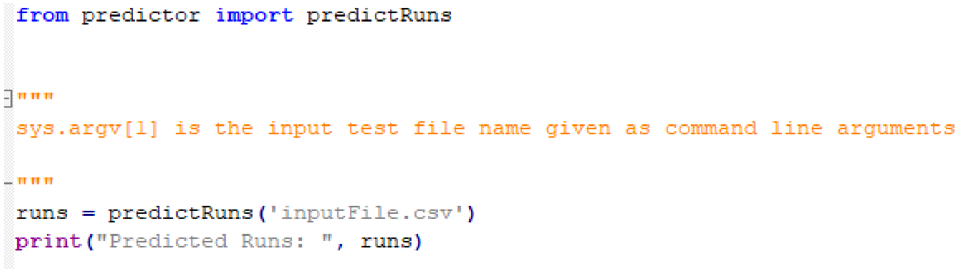
This is the input file:

inputFile.csv:



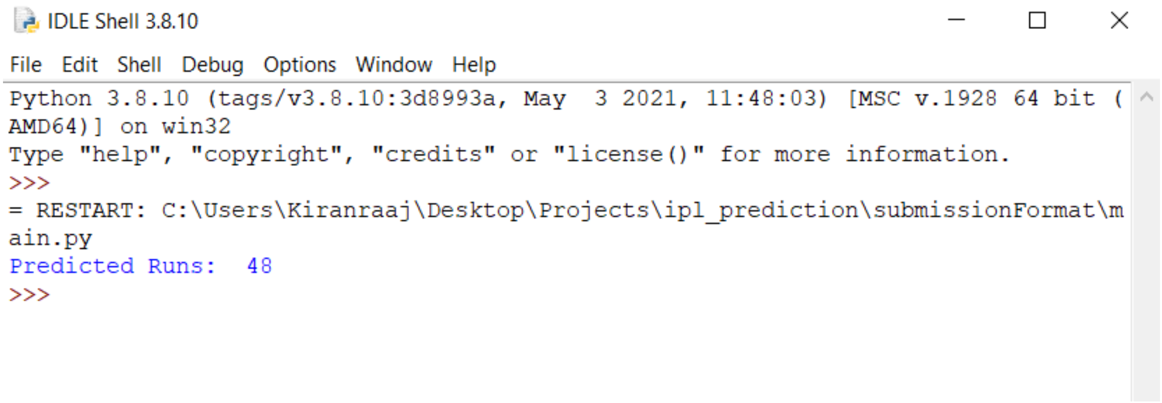
A	B	C	D	E	F
venue	innings	batting_team	bowling_team	batsmen	bowlers
Arun Jaitle	1	Rajasthan Royals	Mumbai Indians	JC Buttler,YBK Jaiswal	TA Boult,J Yadav,JJ Bumrah,NM Coulter-Nile

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:



```
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
>>>
```

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 est 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 cellphones! 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:

R1: Runkler, T. A. (2020). *Data analytics*. Springer Fachmedien Wiesbaden.

R2: Wright, Raymond E. "Logistic regression." (1995).

R3: Marquardt, Donald W., and Ronald D. Snee. "Ridge regression in practice." *The American Statistician* 29.1 (1975): 3-20.

R4: 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.

R5: 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.