2018 International Conference on Computational and Characterization Techniques in Engineering & Sciences (CCTES)
Integral University, Lucknow, India, Sep 14-15, 2018

Data Science – Cosmic Infoset Mining, Modeling and Visualization

Mr. Subhashish Kumar
Dept. of Computer Science and Engineering
Amity University,
Uttar Pradesh, Lucknow, India
officialshubh@outlook.com

Dr. Namrata Dhanda
Dept. of Computer Science and Engineering
Amity University
Uttar Pradesh, Lucknow, India
ndhanda@lko.amity.edu

Mr. Ashutosh Pandey
Dept. of Computer Science and Engineering
Amity University
Uttar Pradesh, Lucknow, India
prncpnd94@gmail.com

***Abstract*— In this paper, we illuminate the research on the statistical analytics expertise field i.e., the Data science, a whole tech world in itself which nowadays has become a buzz word amid geeks. We bestowing the data set mining, modeling and visualization of marking big data with the user friendly open source library of python language by experimenting the real time data sheet focusing on the prophesy job and the technicalities which are necessarily needed by the organization of today's world for the nourishment of future business decision and strategy. Elaborating the keywords provided, from their installation if needed to import and their applicability. As data science is the crux behind the big data analytics and statistics methodology, it has a major role in data field where internet information has a sudden inclination in past four years up to sample of zettabytes and petabytes, where more and more research is needed to make the world parallely excelling in the field of lot of info(s).**

Keywords—Data science, Pandas, seaborn, Numpy, Data mining, Data Visualization, Data sheet Modeling, Education.

I. INTRODUCTION

The vogue of the term "Data Science" has bombarded in technical, academics and business domain, as indicated by a jump in vacancy openings. However, many critical academics and journalists see no distinctions between data science and statistics implementation. Handling with unstructured and structured data, Data Science is a field that encompasses anything related to data cleansing, preparation, and analysis. Data is everywhere and is increasing at an infinite rate. In fact, the amount of digital data that exists is thriving at a rapid rate—in fact, more than 2.7 zettabytes of data exist in today's digital universe, and that is projected to flourish to 180 zettabytes in 2025. That's why more organizations of new world are seeking professional workers who can make sense of all the data. It's the future of development and present for sustainable development. For the sustainability and opulence of data science field, Donoho projects an ever-growing domain for open science where cosmic information sets used for academic publications are accessible to all scholar and industrial researchers. US National Institute of Health has already issued plans to amplify reproducibility and pellucidity of research data. Data science has a conformity that incorporates distinctive degrees of Information, Scientific-Method, Statistics, Advanced Computing, Visualization, Hacker mindset, and Domain Expertise. A professional expertise person of Data Science field is called a Data Scientist. Data Scientists solve obscure and tactical data analysis muddle. The job title has become very noted. On one of the most heavily used employment site, the number of job postings for "data scientist" inclined was more than 100 percent between January 2010 and July 2012. Existence of data scientists helps the companies to make stronger and smarter business decision.

- Amazon prime and Netflix data mines movie interest patterns to analyze on what movie cards a user is interested in, and then uses all the information to predict and generate the movie lists.
- Targets features i.e., what are major range of customer within its base and the unique shopping interests within those group range. This helps them in guiding to message to different market group of audiences.
- Gamble and proctor utilizes time series models to more lucidly and intelligibly stats future need, which helps in planning for optimum production levels.
- Amazon and Flipkart uses recommendation engines for spotting the products, so that it can put the product to remain in the user's vision, using algorithms. Spotify uses algorithms to recommend songs to the user.
- Spam filter of Gmail works with the algorithm for the junk mails and put accordingly the spam, junk and not junk mails in the distinct folders.

Self-driving cars uses computer vision that is also data product- the machine learning code make it able to learn and alert according to the pedestrian, traffic lights and cars on road etc. to obviate accidents. These are the requisites for the professional industrial data scientists.

A. Mathematical Expertise

Mining data and statically analyzing it, is the main challenge for the data scientists to view the data through logical and quantitative oculus. There are several attributes of data such as its delicacy, dimension, and correlation in data that can be expressed graphically with some mathematical applications. Finding panacea by going through the data and making sense of that and predict the next audience target and strategy is bewildering technique. The main solution for the business related problems involve techniques based on hard math, where being able to view and

978-1-5386-4254-2/18/$31.00 ©2018 IEEE

understand intelligently is another mechanism of those method and that is the key to success in building them.

B. *Strong Business Astute*

Data scientist playing major role is expected to be a shrewd, tactical and stalwart business analyzer. Working so hectic with company resources, data scientists are implicated to learn from data in different process, which other can't do. That makes them perfect in observing the data and reflecting it in a graphical or mathematical manner, and contribute to strategy on solving crux business problems. This process establishment makes all the critical points intelligible by data visualization. No data-puking – rather, presenting a very clear shadow of data interpretation and solution, by using data visualization as patronage pillars that lead to guidance.

C. *Technology and Hacking*

At the beginning, it should be cleared that we are *not* talking about hacking as in making the information as key by getting into computers. We're referring to the technical coder subculture meaning of hacking – i.e., creativity and inventive in using technical skills to create or generate things and finding tactical solutions to problems as expressed in Fig. 1.

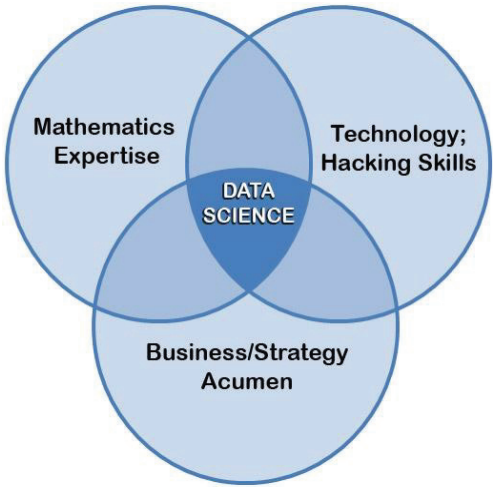

Fig. 1. Technical coder subculture meaning of hacking

II. PANDAS

Pandas is a BSD-licensed, open source library providing efficient and effective-staging, easy to handle data structure, algorithms and data analysis tools for the Python programming language. Pandas is a NumFOCUS sponsored project. The success of development of pandas library is ensured by this library as a world-class open-source project, and make it possible to give it for free to world. Less for the data analyzing and modeling but for big data manipulation and preparation Python has long been great and known. Pandas library provides the function implementation in filling this vacuum, enabling the user to proceed ahead for the data manipulation and visualization without aiming for one other domain like R Programming language.

Aggregated and gathered together with the adroit IPython toolkit and the modules, the domain for working in data implementation in Python make superior in staging, strategic productivity and possibility to aggregate. Pandas does not provide significant data modeling and operating on programs outside of linear and panel regression; for this, look to stats models and scikit-learn. More work is still needed to make Python an outstandingly brilliant class statistical modeling environment, but at present it is well on its way toward the goal.

A. *Installation*

The optimum solution for installing the pandas on system is the command:

conda install pandas

Also can be installed from the PyPI where it has been uploaded using the command:

pip install pandas

B. *Specifications and library highlights*

Tools for writing and reading information between different formats like Microsoft Excel, CSV and text files, SQL databases, and the fast HDF5 format are Intelligent data alignment and integrated handling of missing gaped data: gain automated high label-based alignment in computational techs and easily influence messy data into an orderly and structured manner, staunch label-based slicing, fancy indexing, and sub setting of cosmic data sets, classic performance collaboration and attaching of info sets. Python with pandas is used in a broad and distinct variety of academic and commercial domains, including Finance, Advertising, Web Analytics, Economics, Neuroscience, Statistics and many more.

III. SEABORN

Seaborn is a Python interactive visualization library based on matplotlib. It provides a high-level interface for drawing attractive and user friendly interactive statistical graphics. Many built-in structures for styling matplotlib module's graphics are applicable. Tools exist for choosing color palettes to make different viewing plots that reveal patterns in your information provided. High-level abstractions for manipulating patterns and grids of plots that let you easily build complex modeling and visualization exist. Seaborn performs to aim at production visualization a central part of exploring and understanding data. Seaborn provokes the plot to be user interactive, productive and intelligible with lesser text and more distinct color palettes for better understanding. Matplotlib makes easy things easy and difficult things possible to interpret but seaborn make a well structured set of difficult things easy too. Seaborn provides us customizable graphical formats which lacks in matplotlib. While using open source seaborn library, since available through the pyplot namespace, it is likely that you will often import matplotlib functions to generate easier and less function plots frequently. For the easy installation of latest of seaborn, type on command prompt-

pip install seaborn / conda install seaborn

IV. NUMPY

NumPy or Numerical Python extension is the fundamental package for scientific computing with Python. Used for similar kind of multidimensional array, also used for doing math related stuff or cosmic amount i.e. big data. It is written in C that's why it works fast and glibly. It generally operates on an n-D array i.e. n Dimensional array. It performs functions such as creating an N-dimensional array object, useful linear algebra, Fourier transform, and

random number capabilities, tools for integrating C/C++ and FORTRAN code. NumPy can also be used as an effective and efficient multi-dimensional container of generic data. Arbitrary data-types can be defined. This allows NumPy to seamlessly and speedily integrate with a wide variety of databases.

NumPy is licensed under the BSD license, enabling reuse with few restrictions. Numpy provides the beneficial functions for the numeric games of alteration and visualizing the data set information.

Installation of Numpy goes on by running the code on command prompt – *pip install numpy*.

The basic functioning features of NumPy is its "n-d array", for n-dimensional array, data structure. These n-d arrays are significant process views on memory. Unlike Python's built-in list data structure (which, despite the name, is a dynamic array), these arrays are homogeneously coded: all elements of a single array must be of the same type.

V. DATA SHEET MODELING AND VISUALIZING INFORMATION AND STATISTICS

For the better explanation of modeling and statically visualizing we have taken the original complete data sheet of FIFA championship which includes the past data(s), the columns and rows can be analyzed by running the code-

>> Data_sheet = pandas.read_csv('location://FIFAdataSheet.csv', low_memory = False)

>> Data_sheet.info ()

But before starting, we have to import the libraries needed here so that it can be invoked at the time of operation. The code is as follows-

>> import pandas

>>import seaborn

>>import matplotlib.pyplot

>>import numpy

Hence, it will provide you with the information regarding the whole data sheet with 17981 x 75, rows and columns respectively. For the simplicity we would make a sample data sheet of particular fields that are needed for our analysis and statistics running the command-

>> pre_columns = ['Name', 'Age', 'Photo', 'Nationality', 'Overall', 'Potential', 'Club', 'Value', 'Wage',……., 'Preferred Positions']

>> Data_sheet = pandas.DataFrame(data = fl, columns = pre_columns)

By taking the function help from matplotlib.pyplot and seaborn library which helps to provide user friendly services. Taking the useful data and visualizing it graphically the following queries:

A. *Top ten most potentially strong countries in FIFA championship by running the code-*

>>Top10country = Data_sheet.Nationality.value_counts()

>>a = Top10country.head(10)

>>a.plot(kind = 'pie', legend = True, figsize= (7,7), shadow= True, explode = [0,0.7,0,0,0.3,0,0,0,0,0])

>>matplotlib.pyplot.show()

And the output is shown in Fig. 2.

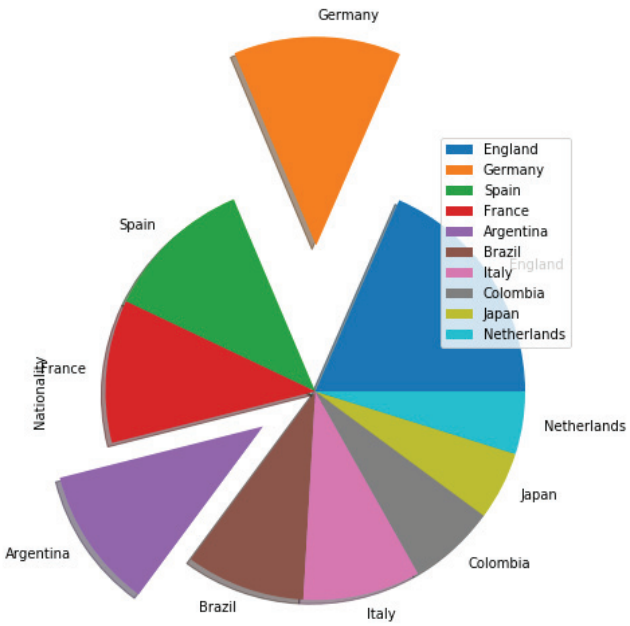

Fig. 2. Top ten most potentially strong countries in FIFA

With this the outcome looks like the Figure 2 representing 'England' team to be the most potential country and 'Netherlands' team to be the tenth in the potential hierarchy. These are the only techniques used in the world class sports match and they predict the data too. Another research came out by the American based Amazon Company, that with these statistics and visualization, it prophecy the order of items of the person, and shows that item in their account in a random manner, and also it is working on data prediction i.e. prophecy of orders and delivery before they are actually ordered by the users. These researches are taking the world on another level of technicalities and accuracy, without any susceptibility, for healthy and sustainable development planning.

B. *Graphical representation on the proportionality of Player's 'Age' and 'Overall Potential'*

Code for the visualization of the proportionality of player's age and overall potential, query is –

>>rbao=Data_sheet.groupby('Age')['Overall'].mean(). reset_index()
>> rbao.sort_values('Age')
>> matplotlib.pyplot.plot(rbao['Age'], rbao['Overall'])
>> matplotlib.pyplot.show()

It will represent the age group having their potential growth and declination. From the data sheet, we found the graph expressed in Fig. 3.

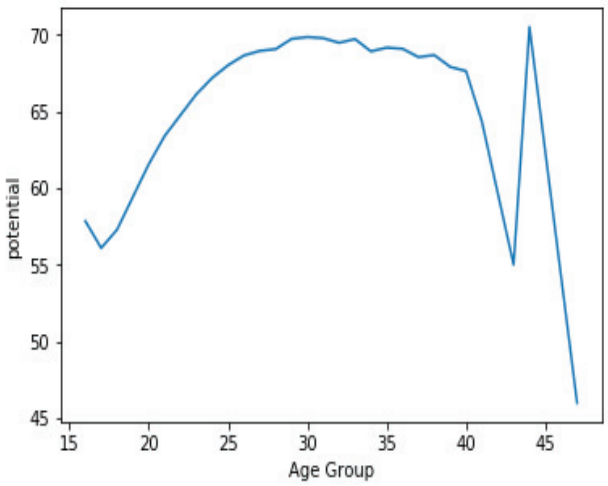

Fig. 3. Graphical representation on the proportionality of Player's 'Age' and 'Overall Potential'

C. *Total FIFA matches played by a country since incite using Seaborn countplot.*

From the ignition year, total number of matches are plotted with the help of python seaborn library titled, 'Nationality' as X- label and 'count' as Y- label can be visualized by the code-

```
>>  matplotlib.pyplot.figure(figsize = (40, 30))
>>  seaborn.countplot(x = 'Nationality', data = Data_sheet)
>>  matplotlib.pyplot.show()
```

The output generated is expressed in the Fig. 4.

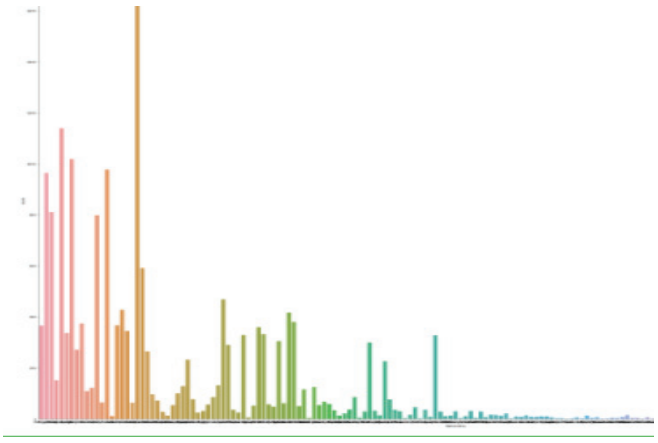

Fig. 4. FIFA matches played by a country since incite using Seaborn countplot

This graph is bestowing the data serially starting from the top i.e. first row and moving towards bottom. The number of played matches fluctuates as they are not structured in sorted manner. Hence, we get a well-represented structured user friendly graph using the palette-I of open source seaborn library.

One more graph representation of the matches played by the teams and 'win by runs' and total matches can be represented as follows in Fig. 5 -

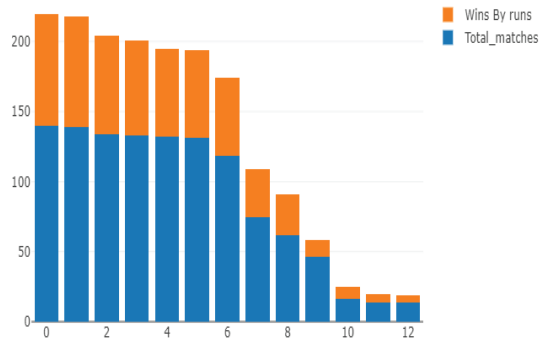

Fig. 5. Graph representation of the matches played by the teams and 'win by runs' and total matches

VI. Conclusion

These are just a sample of what a field of Data science can do, it is beyond our imagination, on which Data scientists and experts are working on and further researches are proceeding which will definitely help the world a lot in celerity. All the cosmic data stored can be used for the general survey and to plan the development and strategies for future purpose. This world full of data is really huge, the thing which matters is how we tackle and strategize it, that's why the demand of data handler or scientists are increasing day by day in the business corporates so make sense of information.

References

[1] March, Salvatore T., and Gerald F. Smith. "Design and natural science research on information technology." Decision support systems 15.4 (1995).

[2] Peffers, Ken, et al. "A design science research methodology for information systems research." Journal of management information systems 24.3 (2007)

[3] https://www.simplilearn.com/data-science-vs-big-data-vs-data-analytics-article.

[4] Baker, Ryan SJD, and Kalina Yacef. "The state of educational data mining in 2009: A review and future visions." JEDM| Journal of Educational Data Mining 1.1 (2009).

[5] https://en.wikipedia.org/wiki/Data_scienceI.S. Jacobs and C.P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.

[6] https://datajobs.com/what-is-data-science

[7] https://pandas.pydata.org/

[8] http://www.numpy.org/

[9] https://seaborn.pydata.org/

[10] https://en.wikipedia.org/wiki/Data_science

[11] https://docs.python.org/3/tutorial/