Turkish Journal of Computer and Mathematics Education Vol.12 No.12 (2021), 3502-3506

Research Article

Intelligent Predictive Modeling Using Big Data for Drug Selection in Pharmaceutical Industry

[1]Netraja Mulay , [2] Dr. Maithili Arjunwadkar

[1, 2] Master In Computer Applications P.E.S.s Modern College of EngineeringPune-05
Pune,India
netraja.mulay@moderncoe.edu.in
Pune,India maithili.arjunwadkar@gmail.com

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 23 May 2021

Abstract—In Pharmaceutical Industries huge amounts of structured, unstructured and semi structured data have been generated. This data is heterogeneous in nature and can be referred to as Big Data. There is a need to manage and analyze this data for taking various decisions regarding drug selection. There are various big data analytical tools and techniques exist with the help of which we can analyze massive amount of data. In this paper we discuss about the various tools and techniques available to analyze the data and selecting the best tool for carting out the predictions regarding right product (drug) selection.

Keywords—Big Data, Analytical Techniques, Drug, R&D

I. INTRODUCTION

A pharmaceutical company is a commercial business whose focus is to research, develop, market and/or distribute drugs, most commonly in the context of healthcare. Indian pharmaceutical industry is now the third largest in the world in terms of volume and stands 14th in terms of value. In Pharmaceutical Industries the Research and Development department is responsible for manufacturing two types of drugs namely Generic drug and Innovator drugs. The Innovator drug is newly invented drug whereas generic drug is a drug which is created from innovator drug when the patent of innovator drug expires then that drug will be called as generic drug. In this paper our focus is on generic drug selection. The generic drugs selling have ooccupied near about 70% to 80% of retail market. These drugs gives rise to the huge volume of data. This data can be generated from various resources like retailers, distributors etc. In order to manage such huge volume or bulky data some big data analytics techniques can be used. Effective analysis of this data will aid the pharmaceutical companies to make the decision regarding right drug selection. ([5] David Asamoah, May 2012)

II. LITERATURE REVIEW

A. Intoduction to Drug Product and Active Pharmaceutical Ingredients

The pharmaceutical company drug production is based on two terms

1. Drug Product(DP)
2. Active Pharmaceutical Ingredients(API)

The drug product is nothing but a drug which can be in the form of tablet, capsules, Injections, syrup, liquid and ointments etc. In order to manufacture a drug product the main component used is Active Pharmaceutical Ingredients (API).

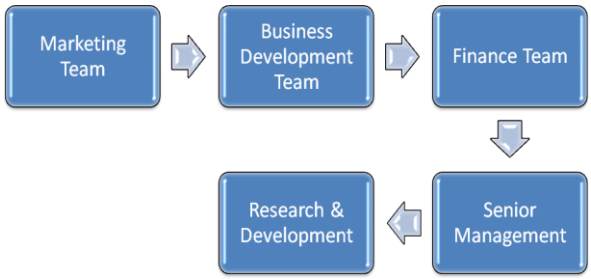
Active Pharmaceutical Ingredients (API) is the active ingredient contained in the medicine. For example, an active ingredient for relief of pain is included in a painkiller. This is called API. A small amount of the active ingredient has an effect, so only a tiny part of the active ingredient is contained in medicine. Some API's for which no drug product is going to be manufactured can be sold out in the market directly. A decision regarding right API and DP selection need to be taken by experts. ([1] Jainul Fathima, January 2018)

B. Drug Development

Drug development is costly process and it requires a lot of time and expenditures also Pharmaceutical industry is data intensive business because it's all vertical generating or using various types of data. The volume of this data is increasing exponentially each day with added variety of sources, so it is very necessary to perform an analysis before undergoing for drug development

C. Mehodolgy for Drug Selection

Every pharmaceutical Industries criteria for drug selection is different but roughly many Industries follow the below mentioned process for drug selection.



- **Marketing Team**- The team of Medical Representatives are acting as a key point of contact between pharmaceutical companies and health care professional. Their role is to identify a demand for an existing product. In order to identify new business opportunity of drugs into the market they interact with different types of people like doctors, nurses, pharmacy owners, hospitals etc. Based on the inputs taken from these people a report is presented to business development team for further process.
  Marketing team gives input regarding the drug to business development team based on the below parameters
  1. **Brand Name**- It refers to the name of the pharmaceutical company currently involved in drug production.
  2. **Volume**-It refers to the volume of production of drugs
  3. **Value**-The current market value of the drug
  4. **Suppliers**-These are the distributors involved in supplying of the drug.
- **Business Development Team**- The major role of business development team is to maintain robust products pipeline and ensure the launch of the right product at right time. The business development team takes into consideration the input generated from marketing team and carry out analysis based on the following points
  1. **Historical data**- It considers the historical data related to that drug. It also takes help of social media in order to gain information regarding the sales trend of that drug.
  2. **Volume and value**- It performs the analysis to determine the volume and value of the drug under consideration.
  3. **Selling Prize**-Selling prize of the drug under consideration is decided by performing its comparison with the drug in the market. This prize must be lower than the prize of the drug in the market.
  4. **Cost of goods sold (Cogs)-**It identifies the actual cost of manufacturing of the drug and it coveys this cost to Research and Development department.
- **Finance Team**- Based on the analysis done by business development team the finance team identifies approximate Investments, Return on Investments (ROI), and payback period of the drug under consideration.
- **Senior Management** – The above analysis done by marketing team, business development team and finance team is put forth in front of senior management team of the company including owner, experts, board of directors etc. After a review senior management recommends the DP and API to be developed to the Research and Development team.
- **Research & Development**- Research and Development department of any pharmaceutical Industry plays a vital role for attainment of company's growth. Introducing a new drug into the market is expensive and risky process therefore proper analysis of the emerging drug selection is very necessary. Any decision taken by this department has their long term consequences and its impact is directly proportional into the market. So based on the inputs received from business development and finance team the Research and Development team carries out following feasibility studies
  1. **Cost of goods sold estimates(Cogs)**-It calculates that whether it is feasible to develop the emerging product or drug in the same cost or lesser than the cost analyzed by business development team.
  2. **Return on Investments (ROI)**-The drug development is expensive and time consuming process. It takes near about 2 to 2.5 years to launch that drug into the market. So before investments we should also think about the Returns on the investments.

Apart from these parameters it also performs analysis regarding number of players, sales trend etc. After successful analysis the drug is filled for the patent approval along with a stability check and then approved drug is taken into consideration for development. ([2]Farzana Elahi)

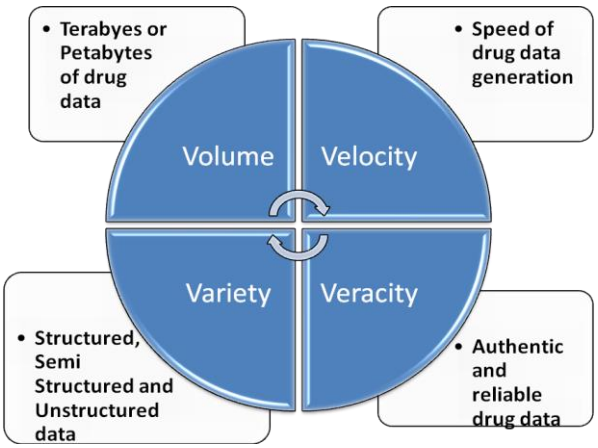
D. Challenges faced by Research and Development Department

Many times because of inadequate knowledge or imperfect market research the drug product may face various challenges. The drugs which are developed at Research and Development Department are rushed to the market without doing any predictions of the market which results into a state of stagnation where in there are only a few products exist in the pipeline. To overcome with the various challenges regarding right product selection the Research and Development department can adopt big data analytics tools and techniques. Use of these predictive analytical techniques helps pharmaceutical companies to select right DP and API, to identify new market opportunities and it will also save time of drug formulation. It also Improves R&D process and help them to release more effective drugs in the market.

III. BIG DATA ANALYTICS

A. Four V's of Big Data

Pharmaceutical company generates the data related to drugs which is huge in nature to manage this data and to carry out analysis regarding drug selection, big data can be used. It refers to a collection of data which is huge in size and always growing exponentially with time. Such data is so large and complex that no any traditional database tool can store or process it efficiently. It is described as holistic and large set information generated from multiple sources at high volume, velocity, veracity and high variety all time. It can be defined using 4 V's



1. **Volume**- It refers to the huge volumes of collected data or a massive scale data which must be managed, analyzed or stored using traditional databases or data processing architecture. The volume of drug data generated by pharmaceutical Industries is growing day by day. The datasets size ranges from terabytes to petabytes beyond.
2. **Velocity**- It refers to the speed of generating the data. In terms of pharmaceutical companies the huge volume of data i.e. stored data should also be processed very fast to gain an insight of the drug instantly.
3. **Variety**- It refers to the different forms of the data i.e. structured, unstructured or semi-structured. The unstructured data includes emails, prescriptions of the doctors, photo, audio, video, hospital medical records etc.
4. **Veracity**-This is the last V which deals with the authenticity or reliability of the data that is being sourced. For example even though social media may not be an authentic data as per present regulations but still that can be used to draw meaningful and important insights specially for new drug development. The various questions of pharmaceutical industries related to drug development like how to capture the sources of data? How to mine the data into usable data? What about storage? How to share these data? How to analyze & present? And most importantly, what insights we can get? The answers to these questions can be given only with the help of big data. Now a day's pharmaceutical industries producing data in mega volumes, even up to terabytes per second. But these datasets may provide information which can provide greater insights regarding product selection. Implementation of Big Data infrastructure enables faster data processing, which, in-turn, allows pharmaceutical companies to support better analytics and derive more focused business outcomes for next-gen research. ([8]Sunil Kumar)

B. Use Of Big Data Analytics in Pharamceutical Industries for DP and API selection

Big data analytics is useful to extract meaningful insight from a large subset of information. It is helpful to the pharmaceutical industries to predict the drug product before development of it. It is an approach to gather, sort and analyze bulky set of data to explore useful insights. Various Big data analytics tools exist to extract voluminous data, analyze data and come arrive at a conclusion or predict regarding selection of a product. The

various data analytics tools are Hadoop, Hive, Mapreduce, Pig and NoSQL Database etc. ([6]Tormay, March 2015)

C. Strength of Predictive Analytics in Pharmaceutical Company

Predictive analytics uses existing data sets, extrapolating the given information and making predictions for future outcomes and possible future trends. It cannot say exactly what will happen, however it can sometimes give a good indication of what is more likely to happen.

Predictive analytics acts as an aid to pharmaceutical companies by providing better insights of data. This analysis will be useful for the R&D team to analyze various parameters like sales trends, Number of players, volume and value of the product, geographical area selection, sector selection (Hospital or Retail) and gross margin etc. Based on these parameters which are analyzed the R&D department can able to make better predictions regarding the selection of drug product or an API. ([3] Preetanshu Pande, October 2016)

IV. TOOLS AND TECHNOLOGIES AVAILABLE FOR DATA ANALYSIS

To handle the huge volume of data of drug product and API and to come arrive at a conclusion regarding the right product selection the pharmaceutical companies may adopt any one of the following technology for analysis and prediction. To manage unstructured data that does not fit into traditional database various tools are used come of these tools are mentioned below.

A. Hadoop to handle Big Data

Apache Hadoop is an ocean of open source software for distributed and parallelized computing. It is specifically used for analyzing and processing large data sets. It consists of a storage part which is known as Hadoop Distributed File System (HDFS), and a processing part called MapReduce. The Hadoop ecosystem contains different tools that are used to help Hadoop modules and offer that functionality. In particular:

- **HDFS-** It is distributed, scalable, and portable file system written in Java for the Hadoop framework. It is designed for processing big data. It uses write once read many model for files. It consists of two types of node 1. A name node and 2. And multiple data node. A name node is single and manages all the metadata needed to store and retrieve the actual data from the data nodes. Metadata stored in name node and application data stored in data node.
- **MapReduce-** This tool is very powerful tool used in healthcare industry. It consists of two stages 1. A mapping stage 2. And a reducing stage. In the mapping stage a mapping procedure is applied to input data. The reducing phase implemented when we want to count the nodes. In this tool map tasks deal with splitting and mapping of data while reduce tasks shuffle and reduce the data.
- **Hive**-It is a platform used to develop SQL TypeScripts to do MapReduce operations. It enables analysis of large data sets using a language very similar to standard ANSI SQL.It can be used with Google Analytics, predictive modeling and hypothesis testing.
- **Spark**-It was developed in response to limitations in MapReduce. It is used in doing real time interactive analysis. It is an analytics engine for large scale data. It can be run by using different programming languages like Python, R, JAVA and Scala. It provides functionality for data processing and analysis, machine learning, graph processing and structured processing ([4] Verbanac, January 2010).

B. Python for Data Analysis

It is the programming language which can be used widely for data analysis. It has rich set of libraries for data analysis.

These libraries can be utilized for big data frameworks. Some of these libraries are mentioned below

- **NumPy-** It is called as Numerical Python. It is designed for numerical applications. It supports multi dimensional arrays and complex arithmetic operations. It has the feature of interoperability also.
- **SciPy**-The Scientific Python library offers advanced operations such as integration, regression and probability. In order to use SciPy, we need to install NumPy first, as it makes use of the underlying modules. It explores usability of the exported modules nicely.
- **Pandas-** It is another data science library which is fast, powerful, flexible and easy to use. It is open source data analysis and manipulation tool. It helps to organize data based on various parameters depending upon requirements. It has rich set of built in data types like series, frames and panels. It has tabular format of frames which allows addition and deletion operation task and grouping of data in easier manner. It also has three dimensional panel data structure which helps in better visualization of the data type This library is very flexible because it supports multiple data formats including missing data.
- **StatsModels-** This module allows user to perform statistical modeling of the data. This is also used for the purpose of forecasting across various domains. As it supports time series analysis capability so for financial organizations it is easy to maintain stock market information in convenient manner. This module is fast enough for big data sets.

- **Matplotlib -** It represents processed graphical data. With the help of this we can generate various types of graphs like pie cart, histogram, bar chart etc. It offers flexibility in its features.
- **Bokeh-**It plots data using web browser interface. It is also used to plot graphs, labels etc.
- **Plotly-** It provides various API's (Application Program Interface) and it is integrated with web applications. ([7]Akshansh Sharma, May 2020)

C. R Language for Data Analysis

R is a free software programming language available for performing statistical computing, data analysis and data visualizations. It has built-in command line interface, but users may use RStudio also. It has big data analysis capabilities. Some of these packages are:

- **pbdR-(Programming with Big Data in R)**
  It is a series of R packages for data analysis and statistical programming for big data.It uses distributed systems and can handle large data sets.
- **MapReduce library in R-** this package allows users to utilize the MapReduce algorithm in R.
- **SparkR-** It is a tool which can be used as a front end for Apache Spark in R

D. Julia

It's a general purpose programming language and very well suited for data analysis and computational science. It aims at scientific computing, machine learning as well as in data mining. It carries out faster and more efficient data analysis as it supports distributed and parallel computing. It can process terabytes of data.

CONCLUSION

In this paper we have seen the impact of traditional method on drug selection. We have also covered the various big data tools and techniques which can be used to carry out effective data analysis but out of these above mentioned technologies python language will give us fast and effective data analysis. Because of rich set of libraries, functions and various integrated open source data analysis and manipulation tools we can carry out predictions regarding right drug (Drug Product or API) selection.

REFERENCES

1. Jainul Fathima,Murugaboopathi Gurusamy,"A Novel Customized Big Data Analytics Framework for Drug Discovery",vol.7,Issue 1 & 2,January 2018.
2. Farzana Elahi, Afia Muqtadir, Shahriyar Anam &K. Mustafiz," Pharmaceutical Product Selection: Application of AHP" International Journal of Business and Management; Vol. 12 ISSN 1833-3850 Canadian Center of Science and Education
3. Preetanshu Pande,R.Bharadwaj," predictive Modeling of Pharmaceutical unit operations",ResearchGate,October 2016.
4. Donatella Verbanac,"Predictive Methods as a powerful tool in Drug Discovery."ResearchGate,January 2010.
5. David Asamoah,Jonathan Annan,"AHP Approach for Supplier Evaluation and Selection in a Pharmaceutical Manufacturing Firm in Ghana",ResearchGate,May 2012
6. Peter Tormay,"Big Data in Pharmaceutical R&D: Creating a SustainableR&D Engine"March 2015.
7. Akshansh Sharma, Firoj Khan, Deepak Sharma, Dr. Sunil Gupta," Python: The Programming Language of Future" Volume 6Issue 12, ISSN: 2349-6002,May 2020
8. Sunil Kumar,Maninder Singh,"Big Data Analytics for Healthcare Industry:Impact,Applications and tools",Volume 2,ISSN:2096-0654