WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

DATA SCIENCE: OPEN RESEARCH ISSUES AND TOOLS

Prathamesh Pundlik Patil[1], Guide: Asst. Prof. Gauri Ansurkar[2]

[1,2]Keraleeya Samajam's Model College, Khambalpada Road, Thakurli, Dombivli (East), prathampatilofficial@gmail.com

ABSTRACT

Data is an effective part of information. Currently, world entered in the era that deals with the large amount of data. The data which we had was mostly structured and less in size, but nowadays the most of the data that we have is unstructured or semi structured and more in size. For overcome with this problem the data science is needed.

In this paper we can see the basic ideas about the data science. Data science is the field where wecan use the scientific algorithms, methods and techniques to extract knowledge from the structured as well as unstructured data. Data science maintains this storage issue with the help of Hadoop framework. Data science also called as a kind of data analysis that will going to inspect the data. This paper shows the much information relevant to Data Science tools and phases of life cycle that complete the Data Science such as business understanding, data collection, data preparation, exploratory data analysis, modeling, model evaluation, and model deployment.

Keywords: *Data analysis, automated discovery, decision making, model evaluation, etc.*

1. INTRODUCTION

Data Science deals with the extraction of learning from substantial volumes of information which are unorganized or unstructured, which is a continuation of the field of perceptive investigation and information mining, otherwise called information mining and information disclosure. Information science frequently obliges dealing with a precise way to measure of data and composing computation to concentrate bits of knowledge from this information. The field ofdata science uses information for planning, insights, and machine learning for researching issues in different spaces during the analysis for data.

Data Science can be considered in two ways firstly, the integration of data sources and analytical and related data processing methodologies. Secondly and quite fundamentally, arising from the convergence of disciplines. Convergence of disciplines is very beneficial in practice. That is, beneficial in regard to addressing and solving problems, and also in regard to the cooperation yielded by cross-disciplinarily.

Basically, data science is the study of the general extraction of knowledge from set of data. Data scientists can solve complex data problems with the help of employing deep expertise in some scientific discipline. Data science majorly focused with the data mining process that leads to removal of unwanted data from the large given data set. In general, Data mining or "Knowledge Discovery in Databases" is the phenomena of discovering patterns in large data sets with database systems, machine learning, statistics and artificial intelligence. The overall need of a data mining process is to extraction formation from a given data set and transforms it into an understandable structure so that it can be further usable. Automated discovery tools have the potential to analyze the raw data and present the extracted high-level information to the analyst, rather than having the analyst finds it for himself or herself.

Data Science is leading to different definitions, one of the most comprehensive definitions of data Science was recently given by Cao as the formula:

Data science = (statistics + informatics + computing + communication + sociology + management)

|(data + environment + thinking).

2. ISSUE IN DATA SCIENCE

Well, data is a lucrative field to pursue as they are plenty of demand for people with similar skills.In this business arena. data scientists are deemed to possess some superhuman powers as they wade across tones of data and come up with a solution for solving business problems.

There exists no career without any challenges and how can data be an exception to this. In this article, we want to explore the real-time challenges of data science which are based on perspectives from those experts in the field.

Problem Identification:

One of the major concerns in analyzing a problem is to identify it accurately for designing a bettersolution and defining each aspect of it. We have seen data scientists to try mechanical approach by beginning their work on data and tools without getting a clear understanding of the business requirement from the client.

There should be a well-defined workflow before starting off with the analysis of the data. Therefore, as a first step, you need to identify the problem very well to design a proper solution and build a checklist to tick off as you analyze the results.

Accessing the Right Data:

It is vital to approach your hands on the right kind of data for the right analysis which can be a little time consuming as you need to access the data in the most proper format. There might be issues ranging from hidden data and insufficient data volume to less data variety. It is also a kindof challenge to gain permission for accessing the data from various businesses.

Data scientists are expected to manage the data management system and other information integration tools such as Stream analytics software which is used for data filtering and aggregation. The software allows connecting all the external data sources and syncing them in the proper workflow.

Cleansing of Data:

Big data is estimated to be a little expensive for generating more revenue because data cleansing is making troubles to operating expenses. It can be a nightmare for every data scientist to work with the databases which are full of inconsistencies and anomalies as unwanted data leads to unwanted results. Here, they work with tons of data and spend a huge amount of time in sanitizingthe data before analyzing.

Data scientists make use of data governance tools for improving their overall accuracy and data formatting. Addition to this, maintaining a data quality should be everyone's goals and businessesneed to function across the enterprise benefit from good quality data. Bad data can result in a bigenterprise issue.

Lack of Professionals:

It is one of the biggest misconceptions to expect that the data scientists are good at high-end toolsand mechanism. But they too need to have possessed a piece of sound knowledge and gain subjectdepth. Data scientists are considered as bridging the gap between the IT department and top management as domain expertise is required for conveying the needs of the business to the IT department and vice Versa.

To resolve this, data scientists need to get more useful insights from businesses in order to understand the problem and work accordingly by modeling the solutions. They also need to focuson the requirement of the businesses by mastering statistical and technical tools.

The Road Ahead:

In reality, being a data scientist requires the implementation of results by making use of refined data and practical applications. The data world is a difficult and fast challenge. However, a careerin the data industry is not only based on experts but it is based on being an expert who understands how to fit the demands of industries.

SAS:

It is one of those data science tools which are specifically designed for statistical operations with the given data set**.** SAS is closed source proprietary software that can be used by large organizations to analyze data. It uses base SAS programming language for providing statistical modeling. It is widely used by companies and professionals working on reliable commercial software. SAS contains numerous statistical libraries and tools that you as a Data Scientist can use for modeling and organizing their data. While SAS is most reliable and has strong support from the company, it is highly expensive and is only used by bigger industries. Furthermore, there are many packages and libraries in SAS that are not available in the base pack and can require an expensive up gradation.

3. TOOLS USED IN DATA SCIENCE

BigML:

BigML, it is another mostly used Data Science Tool. It provides a fully intractable, cloud-based GUI environment that you can enable for processing Machine Learning Algorithms. BigML gives a standardized software using cloud computing for industry requirements. By that, companies can use Machine Learning algorithms across with various parts of their company. For example, it can use this one software across for sales forecasting, product innovations, and risk analytics. BigML specializes in predictive modeling. It uses a wide variety of Machine Learning algorithms like clustering, classification, time-series forecasting, etc.

BigML provides an easy to use web-interface using Rest APIs and you can create a free accountor a premium account based on your data needs. It allows interactive visualizations of data and provides you with the ability to export visual charts on your mobile or IOT devices.

Furthermore, BigML comes with many automation methods that can help you to automate the tuning of hyper parameter models and even automate the workflow of reusable scripts.

D3.js:

JavaScript is used as a client-side scripting language. D3.js, a JavaScript library allows you to make interactive visualizations on your web-browser. With several APIs of D3.js, you can use several functions to create dynamic visualization and analysis of data in your browser. Another powerful feature of D3.js is the usage of animated transitions. D3.js makes documents dynamic by allowing updates on the client side and actively using the change in data to reflect visualizations on the browser.

You can combine this with CSS to create illustrious and transitory visualizations that will help you to implement customized graphs on web-pages. Overall, it can be a very useful tool for Data Scientists who are working on IOT based devices that require client-side interaction for visualization and data processing.

MATLAB:

MATLAB is a multi-paradigm numerical computing environment for processing mathematical information. It is closed-source software that facilitates matrix functions, algorithmic implementation and statistical modeling of data. MATLAB is most widely used in several different scientific disciplines.

In Data Science, MATLAB is taken as for simulating neural networks and fuzzy logic. Using the MATLAB graphics library, you can develop powerful visualizations. MATLAB is also used in image and signal processing. This makes it a very versatile tool for Data Scientists as they can face all the problems, from data cleaning and analysis to more advanced Deep Learning algorithms.

Furthermore, MATLAB's easy integration for enterprise applications and embedded systems make it an ideal Data Science tool. It also helps in automating various tasks ranging from extraction of data to re-use of scripts for decision making. However, it suffers from the limitationof being closed-source proprietary software.

Excel:

Probably the most widely used Data Analysis tool. Microsoft developed Excel mostly for spreadsheet calculations and today, it is widely used for data processing, visualization, and complex calculations. Excel is a powerful analytical tool for Data Science to analyze the data. While it has been the traditional tool for data analysis, Excel still packs a punch.

Excel comes with various formulae, tables, filters, slicers, etc. You can also create your own custom functions and formulae using Excel. While Excel is not for calculating the huge amount of Data, it is still an ideal choice for creating powerful data visualizations and spreadsheets. You can also connect SQL with Excel and can use it to manipulate and analyze data. A lot of Data Scientists use Excel for data cleaning as it provides an intractable GUI environment to pre- process information easily.

With the release of ToolPak for Microsoft Excel, it is now much easier to compute complex analyzations. However, it still pales in comparison with much more advanced Data Science toolslike SAS. Overall, on a small and non-enterprise level, Excel is an ideal tool for data analysis.

ggplot2:

ggplot2 is an advanced data visualization package for the R programming language. The developers created this tool to replace the native graphics package of R and it uses powerful commands to create illustrious visualizations. It is the most widely used library that Data Scientists use form creating visualizations from analyzed data. Ggplot2 is part of tidyverse, a package in R that is designed for Data Science. One way in which ggplot2 is much better than the rest of the data visualizations is aesthetics. With ggplot2, Data Scientists can create customized visualizations in order to engage in enhanced storytelling. Using ggplot2, you can annotate your data in visualizations, add text labels to data points and boost intractability of your graphs. You can also create various styles of maps such as choropleths, cartograms, hexbins, etc. It is the most used data science tool.

Tableau:

Tableau is data visualization software that is packed with powerful graphics to make interactive visualizations. It is focused on industries working in the field of business intelligence. The most important aspect of Tableau is its ability to interface with databases, spreadsheets, OLAP (Online Analytical Processing) cubes, etc. Along with these features, Tableau has the ability to visualize geographical data and for plotting longitudes and latitudes in maps.

Along with visualizations, you can also use its analytics tool to analyze data. Tableau comes withan active community and you can share your findings on the online platform. While Tableau is enterprise software, it comes with a free version called Tableau Public.

Jupyter:

Project Jupyter is an open-source tool based on IPython for helping developers in making open- source software and experiences interactive computing. Jupyter supports multiple languages likeJulia, Python, and R. It is a web-application tool used for writing live code, visualizations, and presentations. Jupyter is a widely popular tool that is designed to address the requirements of Data Science.

It is an intractable environment through which Data Scientists can perform all of their responsibilities. It is also a powerful tool for storytelling as various presentation features are

present in it. Using Jupyter Notebooks, one can perform data cleaning, statistical computation, visualization and create predictive machine learning models. It is 100% open-source and is, therefore, free of cost. There is an online Jupyter environment called Collaboratory which runs on the cloud and stores the data in Google Drive.

Matplotlib:

Matplotlib is a plotting and visualization library developed for Python. It is the most popular toolfor generating graphs with the analyzed data. It is mainly used for plotting complex graphs using simple lines of code. Using this, one can generate bar plots, histograms, scatterplots etc. Matplotlib has several essential modules. One of the most widely used modules is pyplot. It offers a MATLAB like an interface. Pyplot is also an open-source alternative to MATLAB's graphic modules.

Matplotlib is a preferred tool for data visualizations and is used by Data Scientists over other contemporary tools. As a matter of fact, NASA used Matplotlib for illustrating data visualizations during the landing of Phoenix Spacecraft. It is also an ideal tool for beginners in learning data visualization with Python.

NLTK:

Natural Language Processing has emerged as the most popular field in Data Science. It deals with the development of statistical models that help computers understand human language. These statistical models are part of Machine Learning and through several of its algorithms, are able to assist computers in understanding natural language. Python language comes with a collection of libraries called Natural Language Toolkit (NLTK) developed for this particular purpose only.

NLTK is widely used for various language processing techniques like tokenization, stemming, tagging, parsing and machine learning. It consists of over 100 corpora which are a collection of data for building machine learning models. It has a variety of applications such as Parts of Speech Tagging, Word Segmentation, Machine Translation, Text to Speech Speech Recognition,etc.

Scikit-learn:

Scikit-learn is a library based in Python that is used for implementing Machine LearningAlgorithms. It is simple and easy to implement a tool that is widely used for analysis and data science. It supports a variety of features in Machine Learning such as data preprocessing, classification, regression, clustering, dimensionality reduction, etc.

Scikit-learn make it easy to use complex machine learning algorithms. It is therefore in situationsthat require rapid prototyping and is also an ideal platform to perform research requiring basic Machine Learning. It makes use of several underlying libraries of Python such as SciPy, Numpy,Matplotlib, etc.

TensorFlow:

TensorFlow has become a standard tool for Machine Learning. It is widely used for advanced machine learning algorithms like Deep Learning. Developers named TensorFlow after Tensors which are multidimensional arrays. It is an open-source and ever-evolving toolkit which is known for its performance and high computational abilities. TensorFlow can run on both CPUs and GPUs and has recently emerged on more powerful TPU platforms. This gives it an unprecedented edge in terms of the processing power of advanced machine learning algorithms.

Due to its high processing ability, TensorFlow has a variety of applications such as speech recognition, image classification, drug discovery, image and language generation, etc. For Data Scientists specializing in Machine Learning, Tensorflow is a must know tool.

Weka:

Weka or Waikato Environment is used for Knowledge Analysis is a machine learning software written in Java. It is a collection of various Machine Learning algorithms for data mining. Weka consists of various machine learning tools like classification, clustering, regression, visualizationand data preparation.

It is open-source GUI software that allows easier implementation of machine learning algorithms through an intractable platform. You can understand the functioning of Machine Learning on the data without having to write a line of code. It is ideal for Data Scientists who are beginners in Machine Learning.

4. CONCLUSION

The role of statistics in data science is under-estimated as, e.g., compared to computer science. This yields, in particular, for the areas of data acquisition and enrichment as well as for advanced modeling needed for prediction.

Stimulated by this conclusion, statistician as well-advised to more offensively play role in this modernand well accepted field of data science.

Only combining and complementing mathematical methods and computational algorithms with statistical reasoning, particularly for big data, will lead to scientific results based on suitable approaches. Ultimately,only a balanced interplay of all sciences involved will lead to successful solution in data science.

This article has sought to form the foundation for further study of the specific content of data science education and training, and of business sectoral importance.

REFERENCES:

[1] Fionn Murtagh and Keith Devlin, "The Development of Data Science: Implications for Education, Employment, Research and the Data Revolution for Sustainable Development", June 2018.

[2] Claus Weihs and Katja Ickstadt, "Data Science: the Impact of Statistics", January 2018.

[3] Panagiotis Barlas, Ivor Lanning and Cathal Heavey, "A Survey of Open Source Data Science Tools", February 2015.

[4] Proyag Pal, Triparna Mukharjee and Dr. Asoke Nath, "Chellenges in Data Science: A Comprehensive Study on Application and future Trends", IJARCSMS, Vol.3, Issue 8 , August 2015.