COMPUTATIONAL INTELLIGENCE AND MACHINE LEARNING
e-ISSN: 2582-7464

Computational Intelligence and Machine Learning
Vol-4, Issue-1, April 2023

Big Data Analytics Using Machine Learning Techniques for Prediction on Datasets

Ankit Verma[1*], Hansraj[2]

[1] Computer Science and Engineering (M.D.U), Dronacharya College of Engineering, Gurgaon, Haryana, India
[2] Registrar, Dronacharya College of Engineering, Gurgaon, Haryana, India
*Corresponding Author Email: ankiiver@gmail.com

Abstract

Data analytics is the process of performing scientific and statistical analysis on raw data in order to transform it into information that can be used for gaining knowledge. A recently emerging trend in feature abstraction is the combination of computational techniques and big data analysis. This requires gaining knowledge from trustworthy data sources, being able to digest information quickly, and making accurate predictions about the future. The primary objective of this study is to locate the machine learning strategies that produce the most accurate prediction by utilising the model that has been proposed. The supervised and unsupervised strategies have been implemented in a variety of different ways using the MapReduce methodology; however, the suggested model makes use of the Apache Spark framework in order to compare the many existing methods. In this study, the emphasis is placed on elucidating the characteristics of datasets in order to conduct the most accurate analysis possible using machine learning techniques. For the purpose of conducting an analysis of the data sets, machine learning methods such as linear regression, decision trees, random forests, and gradient boosting tree algorithms are utilised. In light of the findings of this research, it is possible to draw the conclusion that when the Spark framework is applied on top of Machine Learning methods, the efficiency of the model is improved by a factor of seventy percent in comparison to the MapReduce paradigm.

Keywords

Apache Spark Framework, Big Data Analytics, Machine Learning Algorithms, MapReduce Paradigm.

INTRODUCTION

In the process of computation, data is transformed into a form that is suitable for the subsequent steps of processing and analysis. In the actual world, data is extremely large, has a complicated structure, and represents even more complex structures, all of which are beyond the capabilities of the architecture that is already in place. In the modern digital age, data is generated in an ongoing stream from an infinite number of sources, and it is made publicly available. The data that is produced by an unlimited number of sources can be categorised in a variety of ways, including as structured, unstructured, or semi-structured data. As a result, the exponential expansion of digitally stored data paves the way for a plethora of new prospects in data analytics. Data analytics is the process of performing scientific and statistical analysis on raw data in order to transform it into information that can be used for gaining knowledge [1]. Data analytics is a collaborative process that uses data to develop complicated judgments from a variety of perspectives in order to address difficulties that arise in the real world. The purpose of analytics is to collect, store, process, and analyse data in order to apply empirical research methodologies to the process of decision making in the actual world. It can be broken down into several categories, including descriptive, inferential, predictive, and prescriptive analytics [2].

The characteristics of Big Data are depicted as ten Vs, each of which is described in figure 1, which is a representation of the characteristics.

Figure 1. Characteristics of Big Data [3]

Big Data refers to a large-scale data processing system that was built to address the problems caused by a variety of conventional approaches. Privacy becomes a significant problem as a result of the vast volume and quantity of data it contains. Acute security can only be achieved by the utilisation of crucial modelling procedures, which call for analytical processing, efficient storage, and effective retrieval methods. As a consequence of this, a massively parallel programming paradigm known as MapReduce is utilised on a distributed architecture in order to provide a high computational environment for Big Data Analytics. In a perfect world, the MapReduce architecture would have two

important functions, namely Map and Reduce, to process the enormous data structures. Within the context of this paradigm, the data are first divided up into parallel distributed map tasks. The data that is input is processed by each map job, which produces pairs of key-value. The output of each task map is then used as the input for the subsequent reduce process. A final summation output is produced by the MapReduce model after it has been processed via several iterations of the reduction task, such as the shuffle, merge, and sort functions. The quality of the finished system is significantly influenced by the method that is utilised to generate such a combination of information and models. In a nutshell, MapReduce models are fundamentally developed for the purpose of directing the experimental research that is conducted in the analytics field [4].

Techniques of machine learning are taught to learn how to interpret complex datasets in order to make important decisions and to tweak features in order to extract high levels of performance. The trained features that are inherent in the parallel programming framework are directly related to the performance that is anticipated. Test features are derived from the learning models that are put through their paces on the datasets used for training. Spark comes equipped with its very own MLlib library, which is specifically designed for machine learning. It resolves issues with streaming, graph computation, and real-time interactive query processing, amongst a wide range of other data-related issues. MLlib provides assistance by capitalising on the scale and speed that is necessary to construct unique use cases for a variety of analytical models [5].

LITERATURE REVIEW

Healthcare, production, sales, IoT devices, the Web, and organisations create a lot of data due to digitalization [6]. Algorithms learn data patterns. Medical professionals and administrators can make decisions using projections. Machine learning algorithms don't require all dataset attributes. Some attributes don't affect prediction. Ignoring or removing unnecessary features reduces algorithm strain. In this work, LDA and PCA are explored on four common Machine Learning (ML) techniques, Decision Tree Induction, Naive Bayes Classifier and Random Forest Classifier, Support Vector Machine (SVM), utilising the Cardiotocography (CTG) dataset through the UC Irvine Machine Learning Repository. PCA always beats LDA. PCA and LDA don't affect Decision Tree and Random Forest performance. DR and IDS datasets test PCA and LDA. PCA-based ML algorithms perform better on high-dimensional datasets, according to experiments. Low-dimensional datasets perform better without dimensionality reduction.

The paper [7] integrates Big Data and Deep Learning to improve intrusion detection systems. Deep Feed-Forward Neural Network (DNN), Random Forest, and Gradient Boosting Tree identify network traffic datasets (GBT). This work evaluates machine learning models using 5-fold cross validation. Results indicate great accuracy with DNN for binary and multiclass classification on UNSW NB15 dataset, 99.16% for binary and 97.01% for multiclass. GBT classifier obtained 99.99% accuracy for binary classification with the CICIDS2017 dataset, whereas DNN achieved 99.56% accuracy for multiclass classification.

Accurate flight delay prediction is key to a more effective airline business [8]. Recent research employs machine learning to anticipate aircraft delays. Most previous prediction systems are single-route or airport-based. This research analyses factors that may influence flight delays and compares machine learning-based models in generalised flight delay prediction tasks. The suggested random forest-based model improves prediction accuracy (90.2% for binary classification) and overcomes overfitting.

Big data analytics are used to accurately forecast and analyse enormous data sets [9]. They reveal hidden information in massive data sets. This research demonstrates a Cloudera-Hadoop-dependent digital infrastructure for analysing any size or type of information. On the basis of real-time Yahoo Finance information, chosen US stocks are analysed to forecast daily gains. The Apache Hadoop big-data framework handles massive data sets through distributed storage and processing. Utilizing Spark's Machine Training program, everyday gain information from US stocks are divided into test and training data sets to forecast stocks with strong daily gains. Banks need bankruptcy prediction models (BPMs) to verify a company's creditworthiness [10]. A robust model needs lots of data and regular updates. Building tools can't directly process that large data.

According to this study, predictive intelligent systems must be built utilising Machine Learning on current models. Data properties drive Machine Learning predictions. Using existing approaches minimises execution time. This model predicts real-time data best using supervised learning in limited time. This study implements Machine Learning on the Apache Spark framework and quantifies the effects of Time and Space dataset complexity.

METHODOLOGY

The dataset used for this research project includes data on temperature. It consists of the annual temperature, which is calculated from January through December. Additionally, seasonal fluctuations that occur throughout the year are included. January to February, March to May, June to September, and October to December are the different time periods for these. The dataset measured for analysis contains temperature data for 115 years, from 1904 to 2016. This data is used to calculate and make predictions for the upcoming years. On this data set, machine learning algorithms are trained to forecast values for future years.

Taking the instances from Amazon Web Services and placing them on EC2 (Amazon Elastic Compute Cloud) allows for the establishment of an environmental setup across cloud platforms. The Jupyter notebook may be viewed on the virtual machine when Spark and Anaconda have been

installed. This allows for the Machine Learning algorithms to be executed. Orange is a tool that is open source and that interprets the output of machine learning algorithms by using a process called cross validation. The tools that are necessary for the execution are outlined in Table 1.

Table 1. Tools Applicable for Implementation

WORK FLOW	TOOLS	VERSION
Job Distribution	Spark	Spark-2.1.0-bin-hadoop2.7
Prediction	Linear Regression Random Forest Decision Tree Gradient Boosting Tree	Anaconda3-4.3.1
Canvas	Orange	Orange 3.2

Figure 2 depicts the work flow that is involved in the establishment of virtual machine instances on cloud platforms. The first step is to get the data set from wherever it was saved in the cloud. For the purpose of exchanging RSA keys on a cloud infrastructure, the open-source key generation programme known as Putty is utilised during the construction of the cloud environment. Instances are made available on EC2 once they have been approved by the authentication process. On the instance, the necessary tools that will be of benefit to the implementation task are installed in accordance with the requirements. Therefore, the infrastructure required for putting the learning algorithms into practise on the virtual machine is made available.

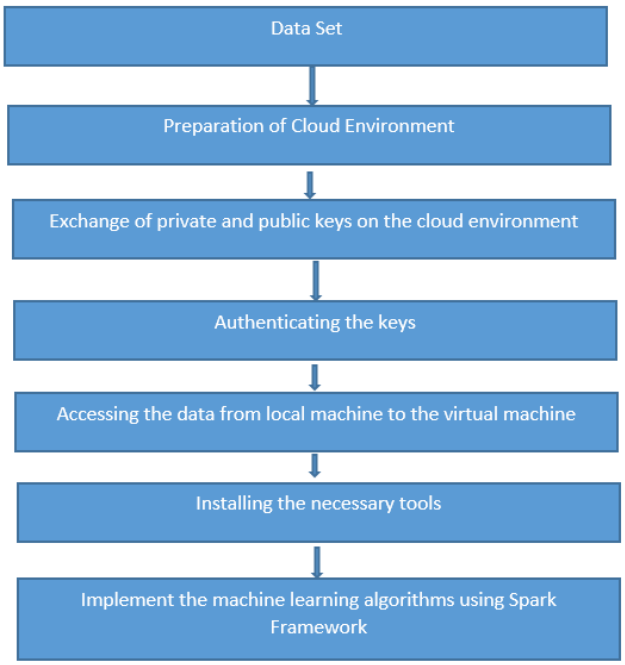

Figure 2. Workflow of Environmental Setup on a Cloud Instance

Three crucial steps make up the suggested paradigm. Data gathering is the primary focus of the first phase, which includes information and a summary of the dataset. Phase Two of the analysis describes the representation of the algorithm, the work flow, and the implementation. Third Phase shows how the learning algorithms' results and errors are represented, with a particular emphasis on comparing the results. The following are the different Machine Learning algorithms that can be used for making predictions: linear regression, random forest, decision tree and gradient boosting tree.

The entities that are shown in figure 3 are incorporated into the proposed system, which is titled "A Framework Model on Big Data Analytics Using Machine Learning Techniques for Prediction on Datasets." This was done, ostensibly, so that these anonymities may be resolved.

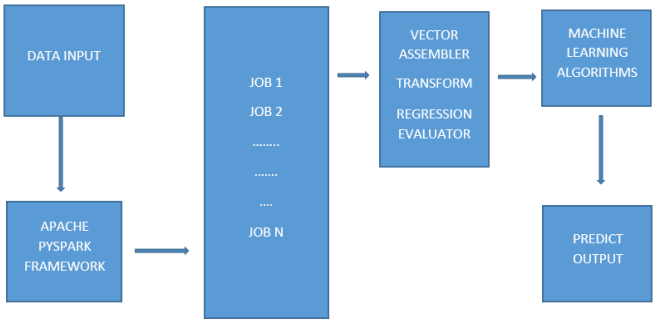

Figure 3. Framework Model for Prediction on Datasets using ML Techniques

RESULTS AND DISCUSSIONS

The study presented here includes a model that may be used to a number of different Machine Learning methods. The graphical representations used here are meant to demonstrate the predictions that were obtained by the learning methods.

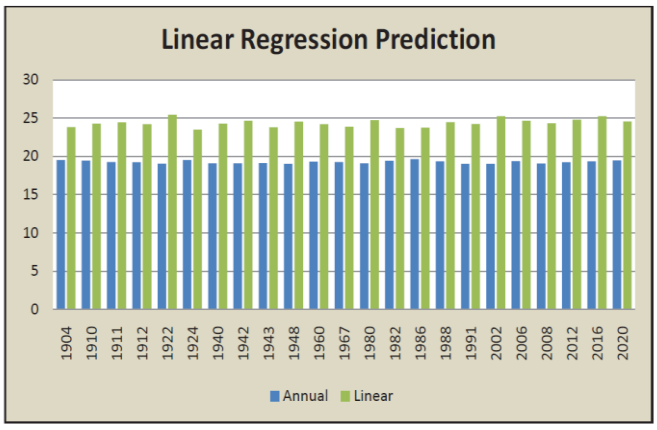

Figure 4. Annual Temperature and Prediction Representation using Linear Regression Algorithm

The linear regression method, which has an accuracy of 72.5%, is used to analyse the annual temperature and compare it to the prediction. Predictions and RMSE are the results of the computations that are drawn from linear

regression techniques. The obtained prediction measures are less precise, and the graphical mapping reveals wide variations; as a consequence, the error rate is higher, despite the fact that time and space were used as efficiently as possible.

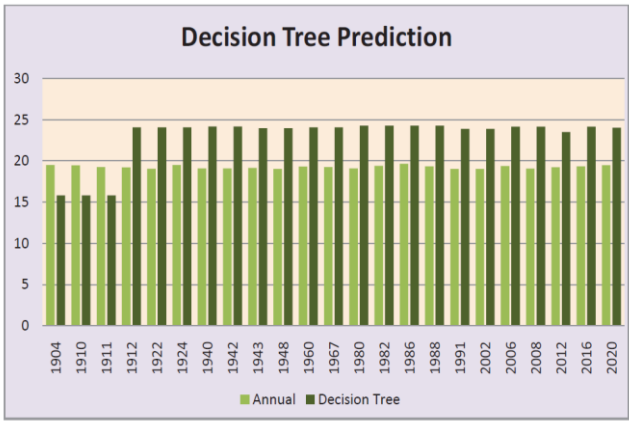

Figure 5. Annual Temperature and Prediction Representation using Decision Tree Algorithm

The Decision Tree Regression approach, which has an accuracy of 82.7%, is used to evaluate the actual annual temperature in comparison to the prediction. Predictions and root mean squared errors are the products of the computations obtained from Decision Tree algorithms. The resulting prediction measures are getting closer and closer to being accurate, and the graphical depiction is showing variances with a rather low error rate. The algorithms take up a significant amount of space.

The annual temperature is measured and compared to the prediction made by Random Forest, which has an accuracy of 95.33 percent. Figure 6 provides an illustration of a random forest that demonstrates relevant prediction to the annual temperature. The results of the random forest prediction method are acquired after generating a forest consisting of twenty decision trees. The ratio of space complexity to overall complexity is higher with the Random Forest approach.

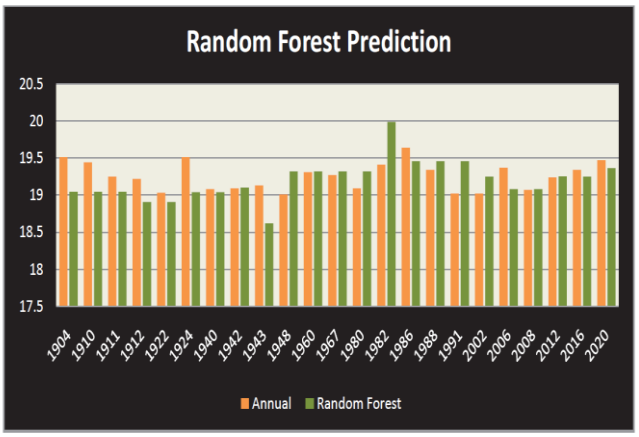

Figure 6. Annual Temperature and Prediction Representation using Random Forest Algorithm

An accuracy of 93.56% is achieved when comparing the actual annual temperature to the prediction made by a Gradient Boosting Tree. Figure 7 provides the tree diagram that summarises the results of the programme.

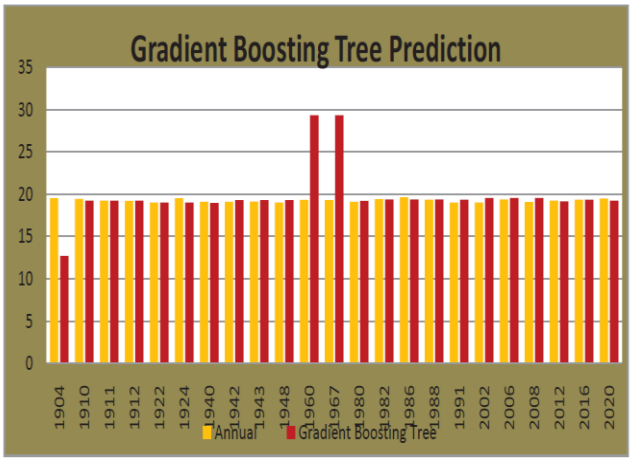

Figure 7. Annual Temperature and Prediction Representation using Gradient Boosting Tree

Prescribed model for prediction uses Linear Regression, Random Forest, Decision Tree, and Gradient Boosting tree techniques. The recorded temperature value reveals Random forest's accuracy. It's more accurate than other ways. Machine Learning approaches have time and space complexity. Machine Learning approaches produce more accurate results. These learning algorithms demand transparency for storage and analysis. For space complexity, linear regression is more efficient than Gradient Boosting Tree. Random Forest predicts more accurately. Decision Tree calculations are accurate (RMSE). This subject extends to cluster task distribution. MapReduce has high read and write disc rates of 28.65 and 35.43 MB/sec, respectively. Reduces processor performance. The Spark platform reads 17.23 MB/Sec and writes 19.41 MB/Sec. Hence Spark improves MapReduce 65.2%.

CONCLUSION

Machine learning approaches are used to observe tasks involving time and space-related aspects while doing data analysis. The suggested approach reduces the overall execution time along with space complexity for upcoming predictions by optimising the machine learning methods in a distributed setting utilising the Spark framework. The model's performance is increased by 70% when Spark framework is added on top of the machine learning methods as compared to that of the MapReduce paradigm. The study provides a comprehensive image of each learning algorithm and their unique properties, and it also examines the presented model using certain metrics.

REFERENCES

[1] Osman, A. M. S. (2019). A novel big data analytics framework for smart cities. *Future Generation Computer Systems*, *91*, 620-633.

[2] Khoshbakht, F., Shiranzaei, A., & Quadri, S. M. K. (2020). ADOPTION OF BIG DATA ANALYTICS FRAMEWORK FOR BUSINESS INTELLIGENCE AND ITS EFFECTIVENESS: AN ANALYSIS. *PalArch's Journal of Archaeology of Egypt/Egyptology*, *17*(9), 4776-4791.

[3] Gupta, S., Kar, A. K., Baabdullah, A., & Al-Khowaiter, W. A. (2018). Big data with cognitive computing: A review for the future. *International Journal of Information Management*, *42*, 78-89.

[4] Mangla, N., & Rathod, P. (2018). Unstructured data analysis and processing using big data tool-hive and machine learning algorithm linear regression. *Int. J. Comput. Eng. Technol*, *9*(2), 61-73.

[5] Mayer, R., & Jacobsen, H. A. (2020). Scalable deep learning on distributed infrastructures: Challenges, techniques, and tools. *ACM Computing Surveys (CSUR)*, *53*(1), 1-37.

[6] Reddy, G. T., Reddy, M. P. K., Lakshmanna, K., Kaluri, R., Rajput, D. S., Srivastava, G., & Baker, T. (2020). Analysis of dimensionality reduction techniques on big data. *IEEE Access*, *8*, 54776-54788.

[7] Faker, O., & Dogdu, E. (2019, April). Intrusion detection using big data and deep learning techniques. In *Proceedings of the 2019 ACM Southeast Conference* (pp. 86-93).

[8] Gui, G., Liu, F., Sun, J., Yang, J., Zhou, Z., & Zhao, D. (2019). Flight delay prediction based on aviation big data and machine learning. *IEEE Transactions on Vehicular Technology*, *69*(1), 140-150.

[9] Peng, Z. (2019, January). Stocks analysis and prediction using big data analytics. In *2019 international conference on intelligent transportation, big data & smart City (ICITBS)* (pp. 309-312). IEEE.

[10] Hafiz, A., Lukumon, O., Muhammad, B., Olugbenga, A., Hakeem, O., & Saheed, A. (2015, March). Bankruptcy prediction of construction businesses: towards a big data analytics approach. In *2015 IEEE first international conference on big data computing service and applications* (pp. 347-352). IEEE.