Journal of Engineering Sciences
Vol 11, Issue 3, MARCH /2020
ISSN NO:0377-9254

CUSTOMER LOAN PREDICTION ANALYSIS

MAMILLAPALLI HARICHANDRA PRASAD* N SRINIVASA RAO**

PG SCHOLAR*, ASSISTANT PROFESSOR**

E-MAIL: harichandraprasad97@gmail.com*, naagaasrinu@gmail.com**

SKBR PG COLLEGE, AMALAPURAM, E.G.DIST, ANDHRA PRADESH – 533201

ABSTRACT:

With the enhancement in the banking sector lots of people are applying for bank loans but the bank has its limited assets which it has to grant to limited people only, so finding out to whom the loan can be granted which will be a safer option for the bank is a typical process. So in this paper we try to reduce this risk factor behind selecting the safe person so as to save lots of bank efforts and assets. This is done by mining the Big Data of the previous records of the people to whom the loan was granted before and on the basis of these records/experiences the machine was trained using the machine learning model which give the most accurate result. The main objective of this paper is to predict whether assigning the loan to particular person will be safe or not. This paper is divided into four sections (i)Data Collection (ii) Comparison of machine learning models on collected data (iii) Training of system on most promising model (iv) Testing.

1. INTRODUCTION

The Iris flower data set or Fisher's Iris data set is a multivariate data set introduced by the British statistician and biologist Ronald Fisher in his 1936 paper The use of multiple measurements in taxonomic problems as an example of linear discriminate analysis. It is sometimes called Anderson's Iris data set because Edgar Anderson collected the data to quantify the morphologic variation of Iris flowers of three related species. Two of the three species were collected in the Gaspé Peninsula "all from the same pasture, and picked on the same day and measured at the same time by the same person with the same apparatus". The data set consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters. Based on the combination of these four features, Fisher developed a linear discriminant model to distinguish the species from each other. The use of this data set in cluster analysis however is not common, since the data set only contains two clusters with rather obvious separation. One of the clusters contains Iris setosa, while the other cluster contains both Iris virginica and Iris versicolor and is not separable without the

species information Fisher used. This makes the data set a good example to explain the difference between supervised and unsupervised techniques in data mining: Fisher's linear discriminant model can only be obtained when the object species are known: class labels and clusters are not necessarily the same.

Nevertheless, all three species of Iris are separable in the projection on the nonlinear branching principal component. The data set is approximated by the closest tree with some penalty for the excessive number of nodes, bending and stretching. Then the so-called "metro map" is constructed. The data points are projected into the closest node. For each node the pie diagram of the projected points is prepared.

The area of the pie is proportional to the number of the projected points. It is clear from the diagram (left) that the absolute majority of the samples of the different Iris species belong to the different nodes. Only a small fraction of Iris-virginica is mixed with Iris-versicolor (the mixed blue-green nodes in the diagram). Therefore, the three species of Iris (Iris setosa, Iris virginica and Iris versicolor) are separable by the unsupervising procedures of nonlinear principal component analysis. To discriminate them, it is sufficient just to select the corresponding nodes on the principal tree.

Existing System:

Machine Learning implementation is a very complex part in terms of Data analytics. Working on the data which deals with prediction and making the code to predicit the future of out comes from the customer is challenging part.

Disadvantages of Existing System:

- Complexity in analyzing the data.
- Prediction is challenging task working in the model
- Coding is complex maintaining multiple methods.
- Libraries support was not that much familiar.

Proposed System:

Python has a is a good area for data analytical which helps us in analyzing the data with better models in data science. The libraries in python makes the predication for loan data and results with multiple terms considering all properties of the customer in terms of predicting.

Vol 11, Issue 3, MARCH /2020
ISSN NO:0377-9254

Advantages:

- Libraries helps to analyse the data.
- Statistical and prediction is very easy comparing to existing technologies.
- Results will be accurate compared to other methodologies.

2. Literature Survey

2.1 BIG DATA

Bata data is a propelling term that depicts any voluminous measure of sorted out, semi-composed and unstructured data that can be burrowed for information. Though huge data doesn't suggest a specific sum, the term is much of the time used when discussing Petabytes and Exabyte's of data.

Bigdata is a term for informational collections that are so extensive or complex that customary information handling application programming is lacking to manage them. Gigantic data is used to depict a tremendous volume of data that is expansive to the point that it's difficult to process. The data is excessively colossal that outperforms current getting ready cutoff. Gigantic Data is an articulation used to mean an enormous volume of both sorted out and unstructured data that is so sweeping it is difficult to process using traditional database and programming frameworks. In most undertaking situations the volume of information is too enormous or it moves too quick or it surpasses current handling limit. Huge Data can possibly enable organizations to enhance operations and make speedier, more keen choices. This information, when caught, organized, controlled, put away, and investigated can enable an organization to increase helpful understanding to expand incomes, to get or hold clients, and enhance operations.

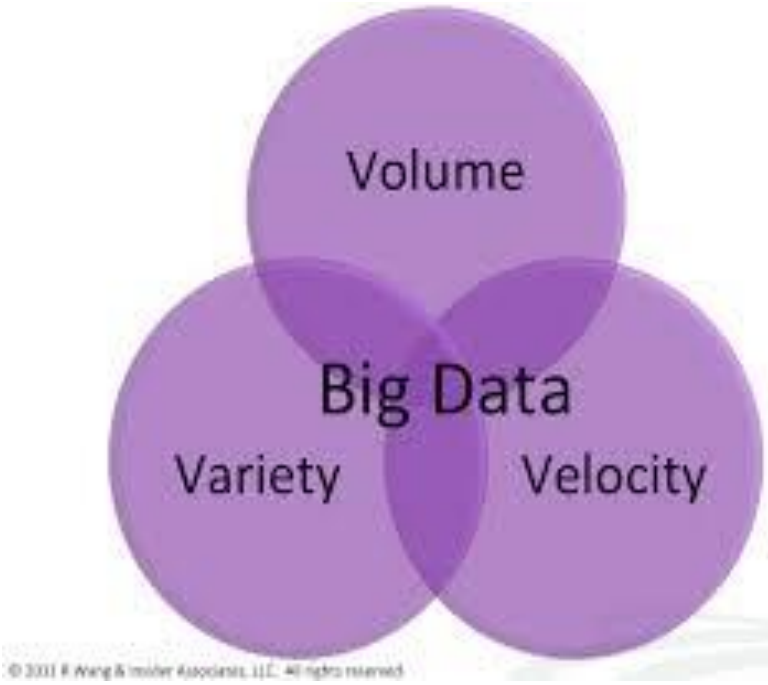

Fig 2.1 Big Data 3V's

Big data can be portrayed by 3Vs: the outrageous volume of information, the wide assortment of sorts of information and the speed at which the information must be must procedures

Volume

Volume is the V most connected with huge information since, well, volume can be huge. Affiliations assemble data from an arrangement of sources, including business trades, web based systems administration and information from sensor or machine-to-machine data. Beforehand, securing it would've been an issue For example, facebook stores pictures of around 250 billions.

Velocity

Speed is the measure of how quick the information is coming in. Information streams in at an extraordinary speed and should be managed in an opportune way. For instance, Facebook needs to deal with a torrent of photos consistently. It needs to ingest everything, process it, document it, and by one means or another, later, have the capacity to recover it.

Variety

Data arrives in an extensive variety of associations – from sorted out, numeric data in standard databases to unstructured substance chronicles, email, video, sound, stock ticker data and cash related trades.

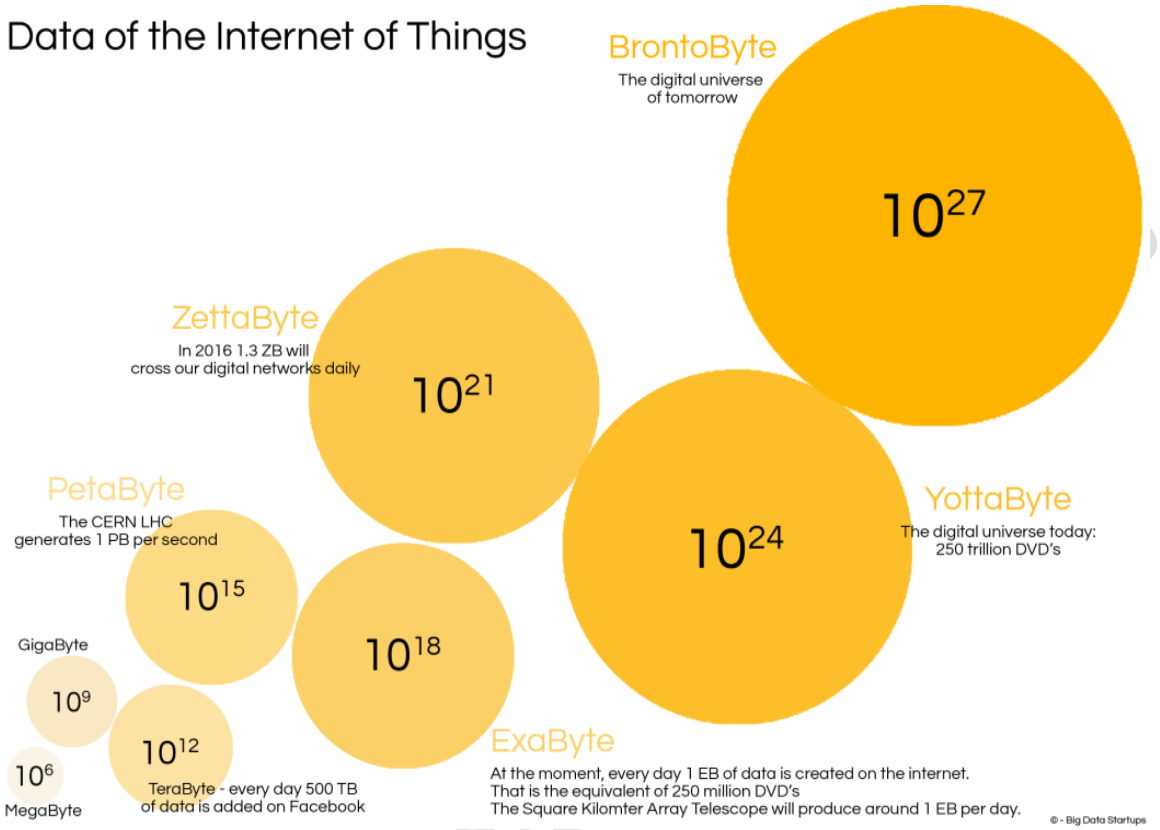

Fig 2.2 Data Measurements

An instance of colossal data might be petabytes (1,024 terabytes) or Exabyte's (1,024 petabytes) of data involving billions to trillions of records

CONCLUSION From a proper analysis of positive points and constraints on the component, it can be safely concluded that the product is a highly efficient component. This application is working properly and meeting to all Banker requirements. This component can be easily plugged in many other systems. There have been numbers cases of computer glitches, errors in content and most important weight of features is fixed in automated prediction system, So in the near future the so –called software could be made more secure, reliable and dynamic weight adjustment .In near future this module of prediction can be integrate with the module of automated processing system. the system is trained on old training dataset in future software can be made such that new testing date should also take part in training data after some fix time.

REFERENCES

[1]. Rattle data mining tool: available from http://rattle.togaware.com/rattle-download.html

[2]. Aafer Y, Du W &Yin H 2013, DroidAPIMiner: 'Mining API-Level Features for Robust Malware Detection in Android', in: Security and privacy in Communication Networks Springer, pp 86-103

[3]. Ekta Gandotra, Divya Bansal, Sanjeev Sofat 2014, 'Malware Analysis and Classification: A Survey'available from http:// www.scirp.org/journal/jis

[4]. K. Hanumantha Rao, G. Srinivas, A. Damodhar, M. Vikas Krishna: Implementation of Anomaly Detection Technique Using Machine Learning Algorithms: Internatinal Journal of Computer Science and Telecommunications (Volume2, Issue3, June 2011).

[5]. J. R. Quinlan. Induction of Decision Tree. Machine Learning, Vol. 1, No. 1. pp. 81-106., 1086.

[6]. Mean Decrease Accuracy https://dinsdalelab.sdsu.edu/metag.stats/code/randomforest.hml

[7]. J.R. Quinlan. Induction of decision trees. MachinelearningSpringer, 1(1):81–106, 1086.

[8]. Andy Liaw and Matthew Wiener. Classification and Regression by randomForest. R News(http://CRAN.R-project.org/doc/Rnews/), 2(3):9–22, 2002.

[9]. S.S. Keerthi and E.G. Gilbert. Convergence of a generalizeSMO algorithm for SVM classifier design. Machine Learning, Springer, 46(1):351–360, 2002.

[10]. J.M. Chambers. Computational methods for data analysis. Applied Statistics, Wiley, 1(2):1–10, 1077