

Credit Analysis in Banking Industry

¹Aniesha Razdan, ²Akshay Naik, ³Mayank Pathak, ⁴Ankush Hutke

^{1,2,3}UG Student, ⁴Assistant Professor, Rajiv Gandhi Institute of Technology, Mumbai, India,

¹ aniesha.razdan11@gmail.com, ² akshaynaik299@gmail.com, ³ mayank74pathak@gmail.com,

⁴ankush.hutke@mctrgit.ac.in

Abstract: Our economy depends on data which is everywhere, in every section, in every country. We produce and convert it into useable data. Information allows us to enhance the business processes and provide our customers and partners with the best quality standards, services and products.

The phenomenal rise in information, and problems encountered while dealing with huge amount of data, make it necessary for an organization to introduce a technique that can overcome these problems and provide an effective solution. Every year the banking organizations, generate enormous amount of valuable data from their customers and their transactions. These valuable data need to be saved and analyzed effectively using big data analytic techniques so as to get the necessary and useful insights for the banking sector.

Big Data Analytics (BDA) provides a better consumer experience with better data management creating transparency, collecting more accurate and detailed performance data, setting up controlling experiments, segmenting populations to customize actions, and replacing/supporting human decisions making with automated algorithms.

The primary focus of our proposed work will be on identifying the issues that the banking sector faces in decision making while granting loans to customers, in detail and providing an optimal solution using the Big Data approach and tools like Hadoop, HDFS, Spark etc.

Keywords — Big Data Analytics, Big data, Hadoop, HDFS, Spark

I. INTRODUCTION

Enhancement in the economic development leads to a huge rise in the requirement of personal loans for customers as the behavior of the borrowers have uncertainty and fuzzy nature. As for the banking industry, banks provide finances to household and the companies by lending loan. One should pay focus towards the riskiness in lending loans to the customers. For both lenders and borrowers, credit risk is a major task, which directly or indirectly affects the reliability of the banks.

Every minute the banking organizations, generate enormous amount of valuable data from their customers and their transactions. These elixir data need to be saved and analyzed effectively using big data analytic techniques so as to get the necessary insights for the banking organizations. In today's market trend, analyzing large amount of data sets comprising of variety of data is of high attention to discover knowledge, market tendencies, customer likings and other business insights. With the capital investment enhancement of many enterprises, credit risk assessment in commercial banks is a crucial area which can have an important impact on stability of bank operation. Therefore, discriminating good borrowers from bad ones with high accuracy is of critical importance to them.

With effective classification among clients, the commercial banks can split customers into different levels by their information, and so the banks can decide whether to provide a loan to their clients according to the classification.

Before approving or rejecting a particular retail loan application, the credit division of a bank thoroughly evaluates that loan application. Loan evaluation refers to utilization of different types of techniques to support automatic decisions that have been utilized so far. Among the models used for this purpose, the decision making process relies on large numbers of historical data spanning over many years of providing credit and decision variables, being statistically analyzed and expressed in crisp values. However, because of information being ambiguous, incomplete, uncertain and imprecise, such approaches cannot help in modeling the way human experts make their decisions about the creditworthiness of the customer.

The banking industry is among many industries which have huge and useful data about their customers but very few banks are utilizing this set of information to enhance the



customer experience and using the data information to prevent intruder. The challenge is not about dealing with trillions of bytes of streaming data, it is about getting initial step with a quantitative approach so that you can drive value from your data, whatever length that data is. They are very well aware of the fact that if the data can be used effectively they can fulfill the needs of customer accurately.

Big Data promises huge impact on the banking and financial services which is why Big Data Analytics is in high demand and becoming essential in making the business decision and providing the biggest edge over the competitors as well as an ocean of opportunities out there for skilled professionals.

Major Scopes of Big Data in Banking and Finance industry in the present and near future include Customer Segmentation, Fraud Detection, Offering Personalized Services, Risk Management, Addressing Compliance Requirements etc. Big data analytics is an emerging trend, particularly in banking and finance and this analytics technology is expected to help the banking industry grow. Banks are taking the big step towards the new need of dealing with the continuous rise in generation of data, with the integration of big data technologies and applications. They have prepared themselves to take in and process large volumes of data created and collected for the future growth of the banking and finance industry, as well as many other business organizations.

II. LITERATURE SURVEY

A lot of research works have been done on ways to enhance decision making for evaluation credit risks while granting loan to potential customers.

Shweta Yadav et al. [1], have measured the credit worthiness of the bank and to calculate the bank performance, they have collected the data and used the big data approach that is Hadoop which is used for analyzing the bank data by performing analytical tests on the data set of the bank based on various parameters so as to know its performance. The analysis has been done so as to know is it worth for a borrower and investors to buy or purchase loan also showing the worthiness of the Hadoop and its ecosystem. Analytical result of the bank loan which reveals the performance and the credit risk of the banks show that there is good opportunity for the borrowers to buy and purchase loan and the investors to invest so as to get higher interest rate with fewer risk but with greater returns.

Chen et al. [2], have studied patterns of default in big data of the banks considering that banks are often worried about the status of repayment because the behaviors of every borrower are always fuzzy and uncertain and ambiguity and credit risk of borrowers must be realized when they review the applications of personal consumer loans. The method of self-organizing mapping (SOM) was applied to split the borrowers into different groups. The result showed 5 groups of borrowers and the 5 degrees of default. Finally, this study comes to know characteristics of borrowers in each groups, and provides the appropriate strategy to each single groups to lower the credit risk in personal consumer loans.

S. Mammadli [3], has proposed a fuzzy logic model for retail loan evaluation. Before approving or denying, especially a retail loan, the credit division of a bank evaluates the loan application. The fuzzy model consists of five input variables such as "income", "credit history", "employment", "character", and "collateral condition" and single output variable which signals credit standing. Whether loan applicant's credit standing shall be considered as "low", "medium" or "high" depends on the control of membership for the linguistic terms of fuzzy output.

C. Xiaojie et al, [4], have utilized a proposed algorithm called simultaneous clustering and attribute discrimination (SCAD) and performs clustering and feature weighting simultaneously. First, the algorithm has been analyzed in detail through a series of compare experiments, confirming this algorithm to have the high clustering precision. Finally, the algorithm is relevant in the analyzing of the bank loan repaid information that can efficiently search the weightage association of the main factors in loan information and realize potential customer.

V. Kumar et al [5], have aimed to analyze the credit risk involved in peer-to-peer (P2P) lending system of "Lending Club" Company. The P2P system allows investors to get worthy higher return on investment as compared to bank deposit, but it comes with a risk of the loan and interest not being repaid. Ensemble machine learning algorithms and preprocessing techniques are used to explore, analyze and determine the factors which play main role in predicting the credit risk involved. A loan is considered "good" if it's repaid with interest and on time. The algorithms are optimized to favor the potential good loans whilst establish defaults or risky credits.

Xudong Lin et al [6], explore the factors affecting the loan willingness of bank and implement some specific measures, which are conducive to solve financing difficult problem for various organizations. They have firstly put forward four crucial factors of which affecting bank's loan willingness, namely target profit, non-performing loan ratio, value customer ratio, business promotion ratio set by the bank itself, then based on likelihood theory, they have measured these factors, and established a simulation model to calculate loan willingness. As a result, with the change of reference point of four main effective factors, and the change of the upper bound and lower bound of four factors, the bank's loan willingness will alter accordingly.

A. Calis et al [7] aimed to smaller the rate of risk to minimum in decision making via analysis of existing personal loan customers and estimate potential customers'



payment outcomes with k-means method, one of the clustering techniques and decision trees method.

IV. EXISTING SYSTEM

The existing system follows a step by step by procedure that takes a lot of time, effort and involves a lot of paperwork. It becomes troublesome for the loan applicants to visit the bank a numerous time to fulfill the formalities and requirements till the application is verified and approved. Also in some cases, because of information being ambiguous, incomplete, uncertain and imprecise, such approaches cannot help in modeling the way human experts

> applicant. Procure the personal loan application form and enter all the required details Submit relevant documents that prove one's income, age, address and identity. The lender will then verify the documents and check the eligibility of the applicant The loan amount will be transferred to the bank account of the applicant if he/she is found eligible.

make their decisions about the creditworthiness of the

Fig 1: Step by step procedure currently followed

The above mentioned step-by-step procedure exhibits the long, time taking process of existing systems that is carried out to grant loan to customers. It takes minimum 7 days for the bank to verify all the credentials of the customer, check his/her eligibility, and evaluate credit risks to ensure that the loan is granted to potential customers with safety assured.

Thus, the main motivation behind this project is to simplify and fasten the process of granting loans to potential customers after validating them thoroughly as well as evaluation of the credit risk involved before sanctioning the loans.

V. RESEARCH METHODOLOGY

The proposed work is used for identifying potential customer after evaluating credit risks for granting loans. Various components of the Hadoop ecosystem have been used to fulfill different stages of the verification process of the loan application. A dataset from Kaggle, world's largest data science community has been used for the implementation. The output given by the model is the decision, whether to grant loan to a particular customer or not.

4.1 Process Flow

Step 1: Initial Phase: Input raw customer data

The customer data of thousands and lakhs of customers associated with the bank is subjected to a cleaning process that includes its pre-processing to transform the raw data into standard and consolidated form. The pre-processing techniques include:

Roszbach et al [8] have proposed a model with a variable

censoring threshold and sample selection effects is for the decision to provide a loan or not and the survival of granted loans. The model is shown to be an intended result tool to separate applicants with short survival times from those with long survivals, the bank's loan provision process is shown to be inefficient. Loans are granted in a way that disagreement with both default risk diminish and survival time maximization. There is thus no trade-off between higher default risk and higher return in the protocol of banks.

Fernando A. F. Ferreira et al [9] provide a study that aims to create a system that integrates cognitive maps and the measuring attractiveness using a categorical based evaluation technique. The multiple criteria expert system delivers results that prove that this approach allows the credit risk evaluation process to be more transparent and informed.

Sudhamathy G. [10] aims to design a model and prototype the same using a data set available in the UCI repository, which is pre-processed and made ready to provide efficient predictions. The idea is to use a decision tree based classification model that uses the functions available in the R Package. The parameters contributing to the prediction and credit risk assessment are identified, and whether the loan applicant is a defaulter or not is predicted.

III. DATA SET

The dataset used has been extracted from Kaggle containing 10,000 entries. This data set contains random data designed manually for enabling users to perform computations on it and draw inferences from it post its detailed analysis. It contains total 18 attributes out of which credit score, annual income and age attributes have been selected for credit analysis.

Attribute Name				
Age	Between 30 and 50 (most preferred)	Numerical		
Credit Score	Credit score type (1: 800 (best score), 2: between 600 and 800 (good score), 3: less than 600 (application rejected))	Numerical		
Annual Income	Higher your income, more will be your scope of borrowing.	Numerical		

Table 1: Dataset Attributes

This dataset was chosen as it provides apt financial information of customers that fulfills the requirement for the analysis to be carried out. As access to actual financial dataset from banks cannot be permitted, this dataset provides the basis for performing analysis using Big Data Analytics.



- Selection of relevant data
- Cleaning to remove errors and inconsistencies
- Integration of data from multiple sources
- Reduction to have a reduced representation of data
- Transformation

Step 2: Application of Big Data Analytics techniques

Using Big Data Analytics techniques, like data mining, regression analysis etc., the data is analyzed in detail to achieve the following and form a training data set.

- Segmentation of customers and Profiling
- Customer Feedback and Analysis
- Fraud Management and Prevention
- Analysis of customer spending patterns
- Risk Assessment etc.

Step 3: Comparison of Customer data with training data set

The training data set will act as a standard against which every customer data and application will be tested, compared to draw conclusions and make decisions effectively.

The analysis will help identify the nature of the customer and transactions made by him/her.

Step 4: Analysis

The application of Data Analytics technique will help in achieving a detailed analysis of customer data and fulfill the objectives:

- Efficient risk management that helps detect errors and frauds
- Correct understanding of customers and their separation from risky ones
- To achieve High precision and accuracy leading to proper data distribution
- Identification of the connection between data captured and possible results
- Transformation of business processes and conducts to identify business opportunities and potential threats.

4.2 Parameters considered for analysis of customer data:

- Credit Score
- Age
- Annual Income

4.3 Conditions for granting loan:

- 1. If Credit_Score > 700, then grant loan.
- 2. If Credit_Score < 600, then

reject application.

 If Credit_Score >= 600 and <=700, then check Annual_Income and age. if Annual_Income > 4,00,000 and age < 55, then grant loan.

VI. RESULTS AND DISCUSSIONS

In this section, we explain the working of our proposed system illustrated with the help of screenshots shown below.

Contractinguist (1) \$ 10000	and an all the second sec	
A R R A COMPANY		
	na pravnik Na pravnik Na pravnik politik politik na politik	

5.1 Beginning with HUE

Hue is a web interface that enables user to browse, query and visualize data. Hue takes the best querying experience with the most intelligent autocompletes, query sharing, result charting and download for any database and helps users find the correct data among thousands of databases and self-document it. The user enters the Hadoop Framework through Hue, where the dataset to be used will be uploaded.

HIND 80	100		m.d. 2 - promising and
848 1	B-factorian		
		Bins and a second balance of the large balance	

Fig 2: Hadoop User Interface

A Designation of the local division of the l	Approximate and the second sec	
A A DE ORIGINAL		
Radia interint Store (b	a de la companya de l Nacional de la companya de la company Companya de la companya de la company Companya de la companya de la company	

5.2 Data stored in HDFS

The Hadoop Distributed File System (HDFS) is the key data storage system used by Hadoop applications. It is a java



based file system that provides expandable, fault tolerance, reliable and cost efficient data storage for Big data. and runs on commodity hardware. The dataset uploaded has to be

stored	in	HDFS	for	processing.
And Advantages of the owner owne		(1.1.1) (1.1.1)		
			t for	
				Automatica des anteses
a filmerer and				- annu <u>- 20</u> 0

Fig 3: Hadoop Distributed File System

5.3 Spark User Interface

Spark is a lightning-fast cluster computing technology, designed for enhance computations which include interactive queries and stream processing, it extends the MapReduce model to efficiently use it for more types of computations, which includes interactive queries and stream processing. The processing starts in Spark using Python programming.

a		
1.1.2 Xolumn		4. 4. O /
Joseffi		
and a shift of the second		
The second second second	entropy and the second	
the state of the state of the state of the	To Tananaki and I	
and its country, Spinster I pass, married them	Control and a state of the stat	Contract to serve the providence of the server of the serv
CONTRACTOR AND DATA AND DESIGN AND ADDRESS.	Verminist Statistics	
and pre-state said sold pre-tablement and the	je state nervel jezetni toža sel v tužna vlast	
The second second second		
to be a subscreen to the second of	Contraction and a local sector of the sector	
the Relation Despired and many lock	Charl drawed with the provide convert the data that they be presented as the	1.2.2.2.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.
reality which has a president the president	TERRETARE AND A CONTRACT OF	
	g a marta finisi tentaku siseu sis	
<u>j stantitisaba</u>	ACCOUNTS AND A COUNTS AND A COU	
		A REAL PROPERTY OF A REAL PROPER

Fig 4: Spark UI

5.4 Data cleaning process by creating Resilient Distributed Dataset

Resilient Distributed Datasets (RDD) is a foundation in data structure of Spark. It is an unchangeable distributed collection of objects. Each dataset in RDD is split into logical partitions, which may be computed on different nodes of the cluster. RDDs can hold any type of Python, Java, or Scala objects, including user-defined classes.

Formally, an RDD is a read-only, distributed collection of records. RDDs can be generated through deterministic operations on either data on stable storage or other RDDs. RDD is a fault-tolerant collection of elements that can be control on in parallel.

Carlos a D and					۰.
				*	
	and the second s				
an an an an an the second s		 ا المحمولين			
(1) STATE CONTRACTOR CONTRACTOR THEY AND ADDRESS CONTRACTOR CONTRACTON CON					
- age in the ball store (when)	0 -	 			

5.5 Hadoop Output: Spark execution using file uploaded on HDFS and view of output in part-00000 file.

The file uploaded on HDFS can be processed using Spark. The previously used RDD creates logical partitions of the dataset. The output files are by default named part-yyyyy where: yyyyy is the task number (zero based).

Contractivities (1) Browning and (1) B (2)			-	۰.	
A A R A MANAGEMENT				0	Ē
and a second sec					
and here he for a manufactor of the property in the second state of the second state o					
fant i dien. I maardatie prij sammerie en prij Armenie in die see aandere	Manual Accession of the second s				
services of an antibility and spectrum and service and the service of the service of the	and and the second s				
the property of the second	Party straight increase, and had a reading source	Contract of the local division of	-		
and the second second rate was seen in the second	Service repaire inclusions	entri kladation, hereidide	-	-	
And the second	A service in a service in a service in the	A REAL PROPERTY OF A LOS OF	-		
	The same lines data incompany, with \$1.45.4	Provide a second second second second second	÷		
	a characterization approximation define the process in	or, starter, which is play. A set-appropri-		_	
in the second size of the second state in the second second second second second second second second second se			-		
and the second	O percendent thesheets (POCCAT)	LABORDER PERMIT			
incomparing and any access to the contract of the second sec					
reprinted a second factors, and integration and interest and to be a lot of the equation of	and participants have been a second of				
marrana acousti a g					
			-		
					67
C P Santana and C C C C C C C C C C C C C C C C C C C					

Fig 6: Hadoop Output

5.6 Loading data into Spark DataFrame

Spark DataFrame is a distributed collection of data organized into named columns that provides operations to filter, group, or compute aggregates, and can be used with Spark SQL. DataFrames can be constructed from existing RDDs, structured data files, tables in Hive, or external databases. It eases the processing of tabular data and enables structured data manipulation.

	a a series	in participa 👔					- 1
1 1 1 1 14 14 14 14 14 14 14 14 14 14 14	-						
Boerda .	10.000.0		a lotter (that is total at	1.0.000, 100, 2010	101111	121	
NIKANGOR	r, 14, 1914 10		Statistics States		the state of the		
		dimension laws	with attention	(b) seen too include	an Intelligence	will come	0.6
and party light in		Annual local	Taxi and	Line and the strength	and an and an an	10000-00	10.0
ten in r. Time .				Control 1 Tanada	and in the last last last	480.00	10.0
The state of the second		P	the state	- 0.0 million - 1	1 Martin Statistics	artist.	10.0
eren ine est. Commit de		minute and	and Distant	1000 A 3		-	10.00
		increased in the	Annaly Constant	-men	A NA COMPANY		10.00
			the state	1000 No. 10	the second of		March .
M.			and in the local division of the local divis	and a second sec	A designation of the local sector of the local	1244.44	1.6
eren er en lige oan e							
erne er en literate e	1. Sec. 1	menter una han	hereast -		and the contract		
content and a Personal de	1.40. () 1.40. ()	many set of		Tipera) a land			***

Fig 7: Loading Data into Spark DataFrame

5.7 Schema of DataFrame in tree format

A schema gives the description of the structure of data. It can be implicit and inferred at runtime, or explicit and



known at compile time. The following showcases schema of DataFrame in tree format.

(4) Add. Aller's brail, spectration and itself (4) A fight (spectration shift) on agent).	Card and a state of the state o			
accelerate Land Land and Land	Contraction of the second second	0.0000000000000000000000000000000000000		1
out the dature Schwart	true must fauther	Concentration of Street of Street, of	parted parties here that he happend	and delivery of restors.
and the second	تعار اللادران ستطريتهمكل	المحموط بتنعد ترجب السالمه	bides, Sol, 21263 Reliage had be they ogt	144
			and the second se	
NAME AND ADDRESS OF AD	2009 204 2011	Her Steller Land	These and part in the second second	1041.11
a the second contract of a local	International Internation	444 (March)	- second contract of the contr	
Canal Annual Committee (Long danis have	And Address of the Ad	I total day to have been been been been	1704-70
A PARTY OF THE PAR	181	P1 100000	- 195,84 M 81,840	1994/198
CONTRACTOR OF ALL CONTRACTOR O	- Long Territ	ALL COMPANY.		1000-001
Party of the second of the	charter one house	and comment	1000 11 0110 1	and an and a second
mai hit	100 C	** 184000		Contraction of the local distance of the loc
AND AND AND ADDRESS OF AN ADDRESS AND ADDR	Jones and Lot	ally depict to the second	Link and the second state	14405-201
anter and the last	useda una toral	anave:	A year at the data to a second s	1846-0461
AND THE STORE STOR	100 C	20 Barrow		ine-silia i
NOT ME COMMENSATION	comment and local	AN DESCRIPTION		COMPANY OF COMPANY
And the set of a set of a set of the set	canner may been	and connect	all page (the last mouthful and	100000
2.4 March and and and and	in the owner where the	ALL ADDRESS	July at the second second second	arrent .
1.41 . 44		at vanet.		and the local division of the local division
		4.11		
and the late in the				
Portion and more and a second	- Properties			
		setting the second second	concerned by the second state to reach the second statement	
	the second second	الإجازية بجيجان الواصف التعاد	Cold Cold State Cold State	weil \$ 1, 17\$ (Total 1, 14 (Total) 1 &
succession was not used on the second second of the		and the second second	Anderer linte Constill Approaphicated for values right	144
		constraint of prints of the state	and the second sec	the state of the s
and the second s	14480 8441 1041	and passed advanced		
				the second se
	and the second s	All and a second se	the second s	
A then been to see the				

Fig 8: Schema of DataFrame in tree format

5.8 Filtering rows from DataFrame based on condition: 'Credit_Score>700'

The dataset can be manipulated, and selected data can be extracted using appropriate queries. Here, the rows from DataFrame are filtered on the basis of the given condition.

marian and the life									
				1000					
All other (party laws ()			-						
where the state of the second state of the sec	cost, ton, manufe ber, ben		will Annual and the	and American	And the second second	et innig and have		inter -	
Los, Springer, 201, 1986, 1991, 1997, 1997, 1997	tring on pasts drong, ap	10.00						_	
the second second processing and the	11.11.11.11.11.11.11.11.11.11.11.11.11.					1			
and the second se									
the second second second second second	in the local division of the local division of the		and the second states of the	Control and Control of	And in case of the	international states			
	COURSES AND		had a second difference	A LOCAL DESIGNATION					
the country in second last and it	And a second second		-	-	to the section of				
81. ····	11	- 12 -	101000	-		T SU - management of			
Aller and the	second lines from	1.00	distanti anno 1		No looked by 1	- Antonio - Cantonio -			
and the second second second in	Company in the second	1.24	deserved in the second se	b comments		and the second second			
41 480	-181	1.16				 It is replaced. 			
Contraction and Contraction and Contraction	Construction (Section)	1400				A. TO Data Labor			
the state of the state of the state of	Second Hard Property	140	1000	and service lines.	torquicket in	and the second s			
-80	144	10	1000	and the second second		an and international			
COLUMN TWO IS AN ADDRESS OF THE OWNER	Sector income	1944		in Later 1	ADD DATE IN				
		10.00				to be immediated			
the second second second	Contract of the second	- 2-	1	and the second second		A DE COMPANY			
time then one - company of the set - 1	The second second second	1.00	-	A local data	manufacture in				
		14		1000		AL BR. LOUDING			
and an are seen as a second set of the set	Manual Annal Street	164	1.000 million (F galanti -	in the state of the				
40 March 199	141	. 86	12,000			0.00 . and .			
ALL AND ADDRESS OF ADD	the second second second	174	11000	- 1- march					
The second					Contra Maria				
and a back of the second									
clearly re- m-ine									
		_					_	_	

Fig 9: Filtering rows from DataFrame

5.9 Filtering rows from DataFrame based on condition: 'Credit_Score<600'

1 11 4 bits manual to					
and and the second second second	100 M 100 M 100	and a second	A S. M. Manufacture	4404.00	
16. T		tree. 1944	 A Westminist 	10.044-141	4.0
The second states		1000 Lower	and the internation	04444	24.07
		inter The	E . S. H. Manner	220.00	24.00
summing into \$4 - taxe					
Orient free houses an inter house					
(and) them defined a		the first has a second before	harris harris	100,000 (top (an and so have
فعالات وارتزعها الرعادة المعارك أراري	cipiel/metricilecto	And the second second	fellower fel.		
Catholic attendances and		والمرابق والمشار المستنا			
and the second s	2000 Con 100 - 200	And Address of the owned	 Becall descentes J. R. Talebar 		24.41
41 41 41 41 41 41 41 41 41 41 41 41 41 4		there to be	and and another	186.41	4.01
an in the set of the set of the set of the	1048-104 Sec. (8)	titeler interior	a hergelige finantigeter	11000.00	14.01
the star site of the second se	COMPANY LONG MADE 100		100 CT	100115	24-01
the last tit. The set of the	alless of the local states			(10.0.01)	14.44
1 10 00 00 00 00 00 00 00 00 00 00 00 00	THE OWNER AND A DESCRIPTION OF A DESCRIP		Construction concentration	(1003-21)	
1. m		Contract of the local division of the local	manufact surgerings of	10010-012	04-81
the last of the state of the state of the	PUBLIC using Tarabi-	constraint in a part of	married (perchaster)	0.000.010	4.0
tere place and . Surfacebor pro- and	talahati may bolat ing	arrest lotter	burger conclusion	0005-01	74-01
And rise and internet and and a 1	1980 - 199 - 1970 - 1890	Table (March	 3.9(30)-520 Arrow Constraints 	1000-00	14.00
T		14400 14400	5 - 10 migrant		
and a state					

Fig 10: Filtering rows from DataFrame

5.10 Filtering rows from DataFrame based on condition: 'Credit_Score>=600 and Credit_Score<=700'

						E		
				1000	in a state of the	the second s	1010-1010-0	100000000000000000000000000000000000000
		_					ind.	_
	10100-00		man or the			and the local		1.444
iter et it	The set		1000		111 T.C	and the second lines		122
and the same and	100202	Ξ.	700 220				and and an and an and an	22
		*	500 13 th			6 (17) w	Control (1998-14)	
	Territoria de la		100 17 10	100	10 10		Carlos and and	22
1000	Gerne .		100 100	H		intetio	and the second s	72
the state	Hun = =	•	100.00		inexes	100000000		22
	finite and a		100 12 20	= 7			and the second s	32
	Terrar .	*	Anne:	-		F and the local	and in the second second	
-			and the second second	-			-	+ 10 100

Fig 11: Filtering rows from DataFrame

5.11 Filtering rows from DataFrame based on condition: 'Annual_Income>400000 and age<55' for credit score between 600 to 700

A REAL PROPERTY AND ADDRESS OF ADDRESS OF ADDRESS ADDR									-
Contraction of Contract								an their manager	
10.100								Contraction of Con-	
	the second second					-		a contraction of the local division of the l	
All some some some some some some some some	the statement		and the second		10.000	1941		1 m 1 m 1 h	_
and contractions. Call in the second	-main antipart		and here !					and the state of a	
			the second se	-				100 Tel # 21	
contraction and the second states		1.00				- and the second			
	1146. 0-010							232514	
a serie Macoli Mediani. Monoral a series de la marte de	and during		100 100	141		1000			
part a service of the part of	10000		the second second	10.0					
Encomposition and the second of the	1000 1010 100		1.			144		A company of the	
Contract of Contractory of the second state of	Odd Andrew		-	-					Sec. 6.14
And the second s	1000 - C 10 - C 10		ALC: NO	100		- 22			
And the second designment of the second design	and some			-		1000	1.00		
the set of	the second second	-			-			to make the first."	
and the second states of the s	This is a lat		1.144	-			100	114 (Har) # . 1	
and the later destances in the state of the state			1.12	-					
and a second state of the second	Contraction of the local division of the loc							100000000	
Caracterization and the second second	and summer			1.4					
A COMPANY OF A DESCRIPTION OF A DESCRIPT	and the second		and inc.	10000	the second second	1000		1.000	
								·	
and the second se									
							-		
C. Charles L. C. Pagette albump. (** 14) 1	and the second se								
								·	
	- Internet in the local		the second second second						
1 Contract (S. 1 - 60) (207, Col. 199)	and the second second	10000		2.82.72				4	
								1000	
			Lo and the second					1000000	
	the second s								
	14 Juli 1			-				And the second sec	-
	12	-		-			- 21	100 C	11

Fig 12: Filtering rows from DataFrame

5.12 Visualization

R is a language and environment for statistical analyses based on computing and graphics. R is also extremely flexible and easy to use when it comes to creating visualizations. One of its capabilities is to produce benchmark quality plots with minimum codes.

It offers a set of essential part of the functions and libraries to build visualizations and present data. For visualization of the output produced and to analyze the trends and patterns, R Studio is used.

In the following images, data that has been filtered from the DataFrames on the basis of given conditions has been visualized using R.

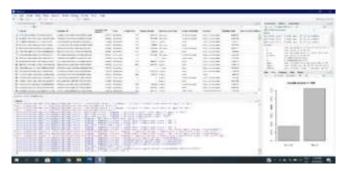


Fig 13: Credit_Score>700



	the last and the last that the last the								
	and a second sec							and the second	and all all all all all all all all all al
	No. in the second secon								
	100 100 100 100 100 100 100 100 100 100	and the set		and the second	di terretaria	Aug. 1	maint in-		
	and the second s	and the second		and have	the second second	-			and the second second second
	and the second se	- indi - inadiran'		and the second	in the second				
	the second secon								the same in the problem
						Sectored Sector	- 846		
								10 March 10	
									and the second s
									the man i w
									And a second sec
									Transmission of the local data
								1.4	
		the second second							
		1000 1000	-						
	Intration in the	CARDING & A DESIGNATION OF						1.1	
	10010000000000000	ALC: NOT						11	
	al and a	dillo						11	
a a a a a a a a a a a a a a a a a a a			110					1	
				1.0.14	-				
				1 2 . I				1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	
				1 2 . I				東京大学	
						CALCULAR CONTRACT		東京大学	
						1998.		軍長王の	-
	Hilli							東京大学	
						INFRE.		東京大学	
						istan.		1 A A A A A A A A A A A A A A A A A A A	

Fig 14: Annual_Income>400000 and age<55 (For credit score between 600 to 700)

From the experimental results, it was found that customers with credit score>600, and of age in between 30 to 50 with high annual income are considered the most by banks while granting loans to the applicants. These customers are categorized as safe and potential customers after evaluating the credit risks involved with respect to each one of them.

VII. CONCLUSION

In this paper, a financial dataset depicting data entries of users has been taken that involves information regarding their loan application. After knowing the issues that the banking sector faces in handling huge amount of data, we chose Big Data Analytics to store, process and analyze the huge amount of data generated particularly when customers apply for loan. To simplify the whole validation and verification process of loan applicants and identify potential candidates, to whom bank can grant loans with minimum credit risk, we proposed our Loan Evaluation model that would overcome drawbacks of the manual process and be beneficial for the customers.

Our proposed work takes into consideration the essential parameters considered by banks for loan purposes including credit score, age and annual income. It verifies the eligibility of the customer based on these parameters and evaluates the credit risks involved while granting loans. This way, it fastens the whole loan application and decision making process by categorizing customers into safe and potential ones to whom the bank can give a nod for granting loan.

REFERENCES

- Shweta Yadav and Sanjeev Thakur, 2017, "Bank Loan Analysis using Customer Usage Data: A Big Data Approach Using Hadoop", 2nd International Conference on Telecommunication and Networks (TEL-NET 2017)
- [2] Chen, Qiao-ling, and Jian-bang Lin, November 2015, "Integrating of business intelligence and CRM in banks: An empirical study of SOM applied in personal customer loans," Fuzzy Theory and Its Applications (iFUZZY), 2015 International Conference on. IEEE.
- [3] S.Mammadli, December 2016, "Fuzzy Logic Based Loan Evaluation System," Procedia Computer Science, pp.495-499.
- [4] C. Xiaojie and D.Huailin, W.Qingfeng, July 2009, "SCAD algorithm and its application in analysis of bank loan," Computer Science & Education, ICCSE'09. 4th International Conference, pp. 1934-1939.

- [5] V. Kumar, Natarajan.S, Keerthana, Chinmayi K.M. and Lakshmi Lupu, September 2012, "Credit Risk Analysis in Peer-to-Peer Lending System" Knowledge Engineering and Applications (ICKEA), IEEE International Conference on IEEE, pp. 193-196.
- [6] Xudong Lin, Lin Cheng, and W.Mao, Z.Qiu, June 2015, "Research on measuring the bank's loan willingness based on prospect theory," Service Systems and Service Management (ICSSSM), 12th International Conference on IEEE, p.p 1-5.
- [7] A.Calis, A.Boyaci, and K.Baynal,March 2015, "Data mining application in banking sector with clustering and classification methods," Industrial Engineering and Operations Management (IEOM),International Conference on IEEE, p.p 1-8.
- [8] Roszbach, Kasper, 1998, "Bank Lending Policy, Credit Scoring and the Survival of Loans," SSE/EFI Working Paper Series in Economics and Finance 261, Stockholm School of Economics.
- [9] Fernando A. F. Ferreira, Ieva Meiduté-Kavaliauskiené, Edmundas K. Zavadskas, Marjan S. Jalali and Sandra M. J, January 2019, "A Judgment-Based Risk Assessment Framework for Consumer Loans," International Journal of Information Technology & Decision Making (IJITDM), World Scientific Publishing Co. Pte. Ltd., vol. 18(01), pages 7-33.
- [10] Sudhamathy G., October 2016, "Credit Risk Analysis and Prediction Modelling of Banks using R", International Journal of Engineering and Technology 8(5):1954-1966.