IRSTI 20.23.17 https://doi.org/10.26577/1i32jpcsit2307

D. Darkenbayev

Al-Farabi Kazakh National University, Almaty, Kazakhstan
e-mail: dauren.kadyrovich@gmail.com

BIG DATA PROCESSING ON THE EXAMPLE OF CREDIT SCORING

Abstract. The article contains the results of research on the development of a model of a Big Data processing system, analysis and forecasting using Data Mining and machine learning methods in solving the problem of mortgage lending, specifically for analysing, forecasting and determining the solvency of individuals receiving a mortgage loan. The article considers the solution of one of the urgent problems of the banking system – mortgage lending. The main problem is to predict the solvency of mortgage borrowers for a long time using the data mining method. The main task is to implement the process of processing big data based on the developed system that determines the solvency of mortgage borrowers. Currently, the pace of long-term mortgage lending is growing annually, as a result, the timeliness of the research carried out in the article on the development of a model of the mortgage lending system, which clearly predicts the solvency of anyone who wants to get housing and makes appropriate decisions, is very relevant.

Key words: development, technology, data, algorithm, analysis.

1. Introduction

The rate of data growth has increased dramatically over the past decade. Research shows that the amount of data has increased approximately tenfold every two years over the past two decades. This overrides Moore's Law, which doubles the processor's performance. About 30,000 gigabytes of data are collected every second and its processing requires high data processing efficiency. Uploading videos, photos, and user posts to social networks collects a large amount of data, including unstructured data. This creates the need to work with large data in various formats, which must be prepared in a special way for further research to obtain modeling and calculation results. In relation to the above, research on big data processing, model development, and solution algorithms is very relevant to this topic. Since the flow of information inevitably increases every year, it is necessary to solve the problems related to the storage and processing of large amounts of data. The relevance of the article is due to the increasing digitization and the increasing transition to virtual technologies. Online activity in various spheres of modern society [1].

The reason for choosing the issue of mortgage loans is that, at the moment, a mortgage loan program is being implemented in the Republic of Kazakhstan, which involves the development of a system for analyzing, determining, and predicting the creditworthiness of mortgage lenders, for a long time; for a long time. Due to the large amount of customer data, these features require a large amount of data to process. In this paper, we develop an efficient mortgage lending software package that uses neural network algorithms to provide models with updated weights over time. This makes the task of data analysis much easier. For example, implementation of software packages developed by financial and credit institutions, quality of service, development of new programs, management, security, etc. This makes things much easier. As the competition between various financial institutions is increasing every year, the need to process customer data immediately and make the right decision in a short time has also increased. Existing tools and systems do not meet the needs of financial institutions. In this context, the research carried out in the paper on big data processing modeling and simulation of the mortgage lending system is very important.

2. The role and importance of Big Data processing in the modern world

Big Data poses a major challenge for improving the management of business processes and transforming information flows into "intelligent" digital resources. There is no strict, universally

accepted definition of the concept of big data. In general, big data refers to the continuous collection of various types of unstructured data [2]. This concept defines large, rapidly growing sets of raw, unstructured data intended for analysis using relational database methods. Terabytes or Petabytes: The exact number is less important than understanding where your data goes and how it's used. The term "big data" was coined by Clifford Lynch, editor-in-chief of Nature, in a September 3, 2008 special issue titled "What technologies can unlock the potential of big data?" Will this affect the future of science? We gather insights from a technology perspective on the rapidly increasing volume and variety of data processed and the potential for a paradigm shift from quantity to quality. [3]

The informative value of Big Data is obvious. Examples of tasks that can be solved with big data flow analysis:

- Predict customer downtime: based on analysis of call center, help desk, and website traffic data.
- Create a predictive model.
- Fraud detection in real-time;
- Risk assessment;
- Build an emergency room.
- Development of operational analysis, etc. [4]

The transition from analog to digital formats accelerates the growth of business information and becomes more serious every day. According to an IDC study, 1 trillion gigabytes of data were generated in 2010, from billions of mobile phones, billions of posts on social media, and an ever-expanding network of sensors used in cars and electric meters. It is used in shipping containers, display cases, commercial terminals, and many other devices [5]. Globally, we understand that data analysis requires large amounts of data processing. In fact, when dealing with unstructured and heterogeneous data, the size of the data does not really matter. To date, there are no universal methods and algorithms for processing and analyzing large amounts of data suitable for all possible cases and situations. If you want to extract insights from raw data, you need to develop algorithms and use different methods for each task. Experts around the world are studying how to process and analyze large amounts of data and how technological advances will affect the future of the economy. All the data collected in the warehouse is not only a hassle to store, but it can also bring significant benefits if properly processed and analyzed. By 2023, the entire global population, including the elderly and children, will have access to 5,200 GB of data. Only 15% of this data is stored in the cloud (as predicted by Lucas Merian, Digital World).

The amount of information doubles every year. IDC estimates that by 2020, 33% of all data will contain information useful for analysis. By 2023, the total amount of information available to humanity will reach 35 zettabytes [6].

3. Existing Big Data storage and processing products

For data collection, processing, and structuring tasks, Oracle Big Data Tools is the solution. It is a pre-installed Hadoop cluster. It is an Oracle NoSQL database and an integration tool with other data stores. Oracle Big Data tools are designed to store and preprocess unstructured or semi-structured data. Which Hadoop-based system works best [7].

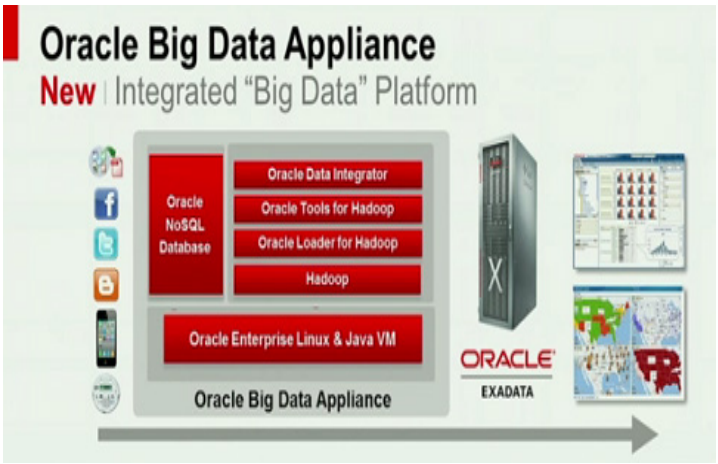


Figure 1 – Data storage scheme in Oracle Big Data Appliance

Gartner analysts have identified three key characteristics of "Big Data" in the so-called "three V".

- Volume: physical size of the stored data;
- speed: speed of data change and subsequent analysis of the results of this change;
- Diversity: different types of data are processed: structured and unstructured data.

The big data workflow model implemented by the Apache Hadoop Project of the Apache Software Foundation (http://hadoop.apache.org) is becoming increasingly popular. Apache Hadoop consists of two components: the Hadoop Distributed File System (HDFS) and the Map Reduce API.

Hadoop is a software platform (software framework) used to create distributed applications for massively parallel processing (MPP). The Hadoop platform consists of two main components.

Hadoop Distributed File System (HDFS) is a distributed file system that provides fast access to application data;

MapReduce is a software platform for distributed processing of large volumes of data on computer clusters [10].

To solve the problem of large amounts of data, a special kind of NoSQL databases (http://www.nosql-database.org) has been developed. A comparison of the properties of relational databases and NoSQL is presented in Table 1.

NoSQL technologies (e.g. Cassandra) are not intended to replace relational databases, but to help solve problems when the amount of data is very large. NoSQL typically uses clusters of low-cost commodity servers. This solution reduces the cost by 1 GB per second several times [6].

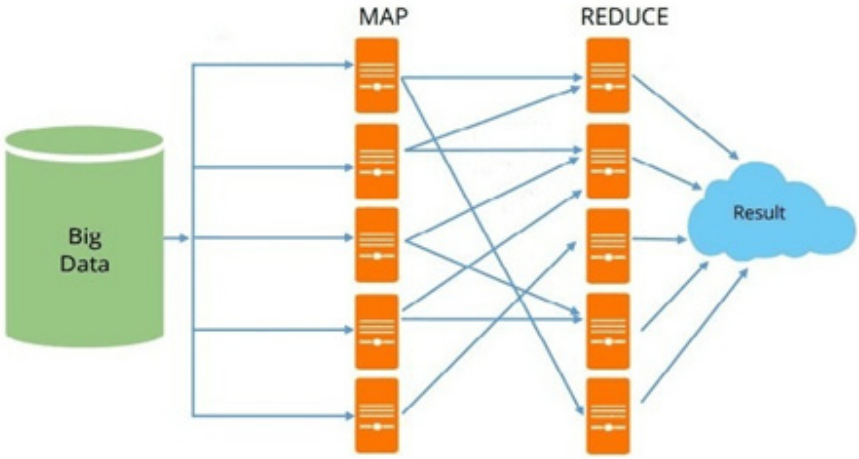


Figure 2 – Scheme of Map-Reduce technology [11]

Table 1 – Comparison of properties of relational databases and NoSQL.

Relational databases	NoSQL databases
Complex Data Relationships	Very simple relationship
Schemacentricity	Arbitrary scheme; unstructured data
Scalability	Distributed processing
Static memory	Memory scales with computing resources
Generic Properties and Functions	The system is focused on the application and the developer

SNA – Shared Nothing Architecture – An independent, distributed computing architecture where each node has its own memory, disk array, and I/O devices. This architecture was developed by IBM in 1974 [12].

It is a programming language (R) for statistical computing and graphics and a free and open-source software environment developed by the GNU Project. R is used everywhere you need to work with data. It is not just statistics in the narrow sense, but also "basic" analysis (graphics, reaction tables) and advanced mathematical modeling. Now, if you are used to using Matlab/Octave programs for professional-level analysis, you can use R without any problem. On the other hand, the core computational power of R is best demonstrated in statistical analysis (from calculators to time series waveforms) [13].

4. Credit scoring as an example of processing large amounts of data

The development of credit markets brings the issue of credit risk to the agenda. In this regard, the creditworthiness of the debtor must be evaluated in order to minimize potential losses.

Changes in macroeconomic indicators and economic, social, and demographic factors can increase the risk of income fluctuation and expose borrowers to the risk of default. This provision involves commercial risks. Large credit groups have similarities, the same product, same insurance, etc. can be grouped into "major loans" based on This group is called a portfolio. The need to group loans into a group. However, such a "big loan" should be characterized by many parameters that allow you to assess the so-called inherent risks. Portfolio risk. The portfolio includes loans that are affected by the same risk factors, both economic (such as industry demand) and social (such as people's income levels). The third level in the hierarchy is credit risk spreads. This risk is the risk arising from the diversification of banking assets according to sectors, focus areas, and banking products. different growth dynamics and different regional economic, industrial, and other conditions. The demand for different types of bank loans determines how diverse the quality of the loan portfolio created by banks is [14].

Credit scoring is a way to evaluate and manage your business credit.

Scoring is a mathematical or statistical model for determining the likelihood of a customer repaying their loan within a certain period of time, based on the credit history of the customer who has previously used the bank's services. Consider the possibility of bankruptcy of potential borrowers when borrowing [15]. As a result of the overall evaluation of the customer's points, the customer will be given a certain number of points, taking into account the delivery date. The main scoring tool is the scorecard. This is a mathematical model that allows you to compare borrower characteristics with a numerical value to get a score. One of the most obvious examples of the use of big data processing is scoring, which forms the basis of risk analysis in banking systems.

The main methods and techniques for the development of evaluation models for banking risk management systems, including algorithms for processing large amounts of data, are banking studies. These challenges are being solved by improving the distribution of application data in banking services creating accurate and appropriate rating models using current technology and developing new algorithms. The automatic nature of the evaluation model plays an important role. A credit score is one of the guides for evaluating potential borrowers, real or legal persons, due diligence, and financial situation before deciding on a loan. The term "Score" comes from the English word "Score" and means achievement, result, amount of debt, cause, reason, etc. This term can be understood in both a broad and narrow sense.

Scoring is a method of dividing the entire test sample into different groups. As shown in Figure 3 for credit products, there are 'bad' and 'good' customer groups. Although we do not know in advance whether a customer will repay the loan, we do know other factors that help us determine which group the customer belongs to. The idea of classifying populations in statistics was developed by Fisher in 1936, using plants as an example.

David Duran first used the same technique in 1941 to classify loans into "bad" and "good" regardless of whether the loans were in default.

In the article, we made a forecast by determining their solvency and deciding whether to give this client a mortgage loan or not. Processing your big data by their attributes. For processing, I used DataMining algorithms: linear regression, logistic regression and multilayer neural network. The expected result is "1" for good clients and "0" for bad clients. If "1" then you can give the client a mortgage loan, and if "0" then no.

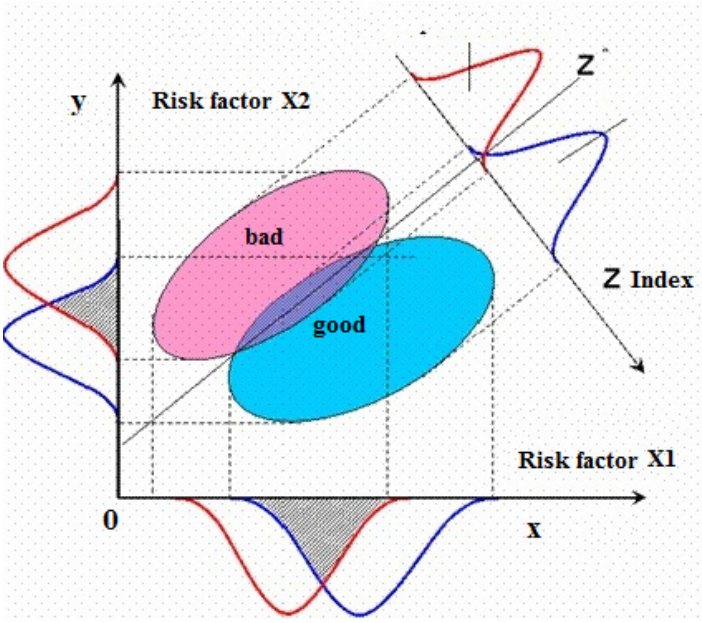


Figure 3 – Geometric interpretation of credit scoring

5. Using a MongoDB database to store big customer data

Working with large amounts of information in the context of databases is a separate problem. As the amount of information increases, you face problems with disk space consumption (solving this part of the problem is relatively simple) and, most importantly, with access to the necessary data over time. You can use advanced caching, but in the end, it will not work for you. You can split the database by placing each category of information in a separate database.

The performance of the system decreases significantly as the database size increases. One way to reduce data access time is to store the database in RAM. This technology enables a 100-fold increase in efficiency [6]. Memory databases – IMDB – databases that use computer RAM to store data; RAM is the main storage medium in such systems. Since the cost of RAM is decreasing day by day, using it as memory makes sense to increase data processing speed.

There are new types of databases for working with large amounts of data, such as databases with built-in analytics. Currently, almost all databases use this concept [16]. However, Teradata developers were the first to develop database-based analytics [17].

Additionally, a database type is a column used to store data. In recent years, many column-oriented database systems have emerged, including Mongo DB, Monet DB [18, 19], and C-Store [20]. The developers of these systems claim that their approach provides large performance gains for certain workloads, particularly analytical workloads that require large numbers of data reads, such as in data warehouse applications [21].

6. Analytical platform

Unlike a DBMS, which includes a set of data extraction algorithms, an analytics platform is designed to focus first on data analysis and create ready-to-use analytical solutions.

An analysis platform is a specialized software solution that includes an information analysis system and all the tools for extracting patterns from “raw data”. A process is performed to extract specific patterns from the entire range of data. This includes means of integrating information into a single source (storage data), extraction, transformation, data transformation, data mining algorithms, visualization, methods, and simple and complex models. To move forward [22].

There is no single method or algorithm that correctly handles large amounts of data. Therefore, each task has its own implementation algorithm and data analysis algorithm. There are many algorithms for processing large amounts of data. However, depending on your purpose, you may need to develop new algorithms “for your own use”.

MongoDB is a document-oriented database management system that does not require table schema details. It is considered one of the classic examples of NoSQL systems and uses JSON-like documents and database schemas. Written in C++.

7. Data preparation

We also created a database in MongoDB. We are transforming the data into MongoDB for data processing.

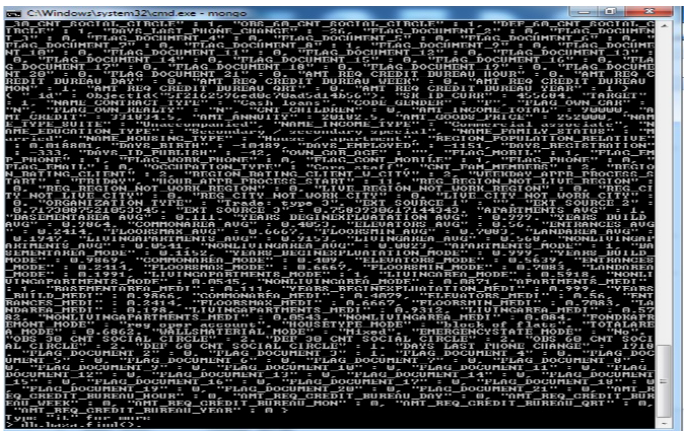
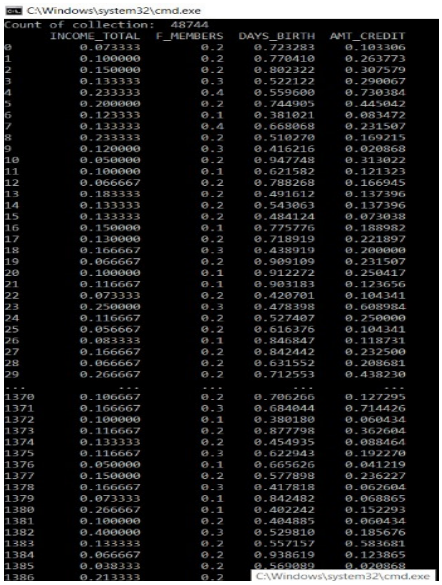


Figure 4 – Converting data to a MongoDB database

8. Data normalization

Normalization is the restoration of data from its original range so that all values are between 0 and 1 [23]. Normalization can be useful for some machine learning algorithms because in some cases the input values of time series data have different sizes. This is necessary for k-nearest neighbor algorithms, as it is used in distance calculation and linear regression, artificial neural network weighting. Normalization requires an accurate estimate of the minimum and maximum values in our calculation. If the time series has an uptrend or downtrend, it is difficult to determine the expected values, and in this case, normalization is not considered the optimal way to solve the problem. The value is normalized as follows[24]:

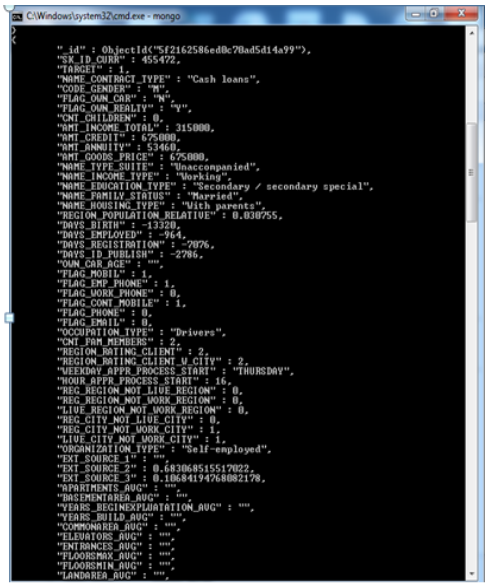
$$y = (x - \min)/(\max - \min) \qquad (1)$$



Count of collection: 48744

	INCOME_TOTAL	F_MEMBERS	DAYS_BIRTH	AMT_CREDIT
0	0.023333	0.2	0.723283	0.103306
1	0.100000	0.2	0.770410	0.263773
2	0.150000	0.2	0.802322	0.307579
3	0.133333	0.3	0.522122	0.290067
4	0.233333	0.4	0.559600	0.730384
5	0.200000	0.2	0.744905	0.445042
6	0.123333	0.1	0.381021	0.083472
7	0.133333	0.4	0.668068	0.231507
8	0.233333	0.2	0.510270	0.169215
9	0.120000	0.3	0.416216	0.020868
10	0.050000	0.2	0.947748	0.313022
11	0.100000	0.1	0.621582	0.121323
12	0.066667	0.2	0.788268	0.166945
13	0.183333	0.2	0.491612	0.137396
14	0.133333	0.2	0.543063	0.137396
15	0.133333	0.2	0.484124	0.073038
16	0.150000	0.1	0.775276	0.188982
17	0.130000	0.2	0.718919	0.221897
18	0.166667	0.3	0.438919	0.200000
19	0.066667	0.2	0.909109	0.231507
20	0.100000	0.1	0.912272	0.250417
21	0.116667	0.1	0.903183	0.123656
22	0.073333	0.2	0.420701	0.104341
23	0.250000	0.3	0.478398	0.608984
24	0.116667	0.2	0.527407	0.250000
25	0.056667	0.2	0.616375	0.104341
26	0.083333	0.1	0.840847	0.118731
27	0.166667	0.2	0.842442	0.232500
28	0.066667	0.2	0.631552	0.208681
29	0.266667	0.2	0.712553	0.438230
...	...	...	...	...
1370	0.106667	0.2	0.706266	0.127095
1371	0.166667	0.3	0.684044	0.714426
1372	0.100000	0.1	0.380180	0.060434
1373	0.116667	0.2	0.877798	0.362604
1374	0.133333	0.2	0.454935	0.088464
1375	0.116667	0.3	0.622943	0.192270
1376	0.050000	0.1	0.665626	0.041219
1377	0.150000	0.2	0.577898	0.236227
1378	0.166667	0.3	0.417818	0.062604
1379	0.073333	0.1	0.842482	0.068865
1380	0.266667	0.1	0.402242	0.152293
1381	0.100000	0.2	0.404885	0.060434
1382	0.400000	0.3	0.529810	0.185676
1383	0.133333	0.2	0.557157	0.583681
1384	0.066667	0.2	0.938619	0.123865
1385	0.038333	0.2	0.569089	0.070869
1386	0.213333	0.2		

Figure 5 – Normalized data




```
{
    "_id" : ObjectId("5f2162586ed0c70ad5d14a99"),
    "SK_ID_CURR" : 455472,
    "TARGET" : 1,
    "NAME_CONTRACT_TYPE" : "Cash loans",
    "CODE_GENDER" : "M",
    "FLAG_OWN_CAR" : "N",
    "FLAG_OWN_REALTY" : "Y",
    "CNT_CHILDREN" : 0,
    "AMT_INCOME_TOTAL" : 315000,
    "AMT_CREDIT" : 675000,
    "AMT_ANNUITY" : 53460,
    "AMT_GOODS_PRICE" : 675000,
    "NAME_TYPE_SUITE" : "Unaccompanied",
    "NAME_INCOME_TYPE" : "Working",
    "NAME_EDUCATION_TYPE" : "Secondary / secondary special",
    "NAME_FAMILY_STATUS" : "Married",
    "NAME_HOUSING_TYPE" : "With parents",
    "REGION_POPULATION_RELATIVE" : 0.030755,
    "DAYS_BIRTH" : -13320,
    "DAYS_EMPLOYED" : -964,
    "DAYS_REGISTRATION" : -7076,
    "DAYS_ID_PUBLISH" : -2786,
    "OWN_CAR_AGE" : "",
    "FLAG_MOBIL" : 1,
    "FLAG_EMP_PHONE" : 1,
    "FLAG_WORK_PHONE" : 0,
    "FLAG_CONT_MOBILE" : 1,
    "FLAG_PHONE" : 0,
    "FLAG_EMAIL" : 0,
    "OCCUPATION_TYPE" : "Drivers",
    "CNT_FAM_MEMBERS" : 2,
    "REGION_RATING_CLIENT" : 2,
    "REGION_RATING_CLIENT_W_CITY" : 2,
    "WEEKDAY_APPR_PROCESS_START" : "THURSDAY",
    "HOUR_APPR_PROCESS_START" : 16,
    "REG_REGION_NOT_LIVE_REGION" : 0,
    "REG_REGION_NOT_WORK_REGION" : 0,
    "LIVE_REGION_NOT_WORK_REGION" : 0,
    "REG_CITY_NOT_LIVE_CITY" : 0,
    "REG_CITY_NOT_WORK_CITY" : 1,
    "LIVE_CITY_NOT_WORK_CITY" : 1,
    "ORGANIZATION_TYPE" : "Self-employed",
    "EXT_SOURCE_1" : "",
    "EXT_SOURCE_2" : 0.6830685515517022,
    "EXT_SOURCE_3" : 0.10684194768082178,
    "APARTMENTS_AVG" : "",
    "BASEMENTAREA_AVG" : "",
    "YEARS_BEGINEXPLUATATION_AVG" : "",
    "YEARS_BUILD_AVG" : "",
    "COMMONAREA_AVG" : "",
    "ELEVATORS_AVG" : "",
    "ENTRANCES_AVG" : "",
    "FLOORSMAX_AVG" : "",
    "FLOORSMIN_AVG" : "",
    "LANDAREA_AVG" : "",
```


Figure 6 – Data from Mongo DB database

9. The principle of operation of the developed data processing model

In the model, the MongoDB database is used to store large data that is not structured. The database function is easily integrated with various programming languages, that is, it is convenient for data transformation, import, data normalization[25].

Here is the data in the form of a matrix, this is the value for training the data we know. – error decreasing with each iteration of the loop, – weighting factor, – epoch. Giving 100 epochs with three machine learning algorithms and manually introducing the error, we get the expected result. The accuracy of the three used algorithms is compared and the algorithm with the best result is selected. Next, the process of processing test data from our MongoDB database is processed by a high-performance algorithm. One of the main tasks in creating a numerical model is to assess the payment capabilities of individuals by processing large amounts of unstructured data. If the MongoDB database is chosen as a single database for mortgage lending organizations, then it is clear that many tasks will be optimally solved. If customer data were immediately retrieved from the database and processed, complex problems with the distribution of apartments would be resolved positively.

Table 2 presents statistics from our MongoDB database to determine the solvency of mortgage borrowers. A total of 356,256 test data were used to test the effectiveness of the system model in determining solvency. Of these, 48,744 people made up the test training sample. In order for us to accurately predict the data, we need to split 100% of the data into 25% of the test and the remaining 75% into training. With the help of general software, 1,425,116 records of 356,279 people were processed.

The set used to model the system for determining the solvency of individuals is presented in table 2 below.

Table 2 – Data sets used for processing.

Data sets	Total	Total
Train	307 511	356 279
Test	48 768	

10. Development of a software package and the results of processing Big Data

The scheme of the workflow of the system for determining the solvency of borrowers of long-term mortgage loans, created in the high-performance Spider environment of the Python software package for computer scientists, is presented[8]. According to it, the user first gets into the main 4 menus of the program, which are based on determining the solvency of the borrower on a long mortgage loan. They are called: “Basic”, “Assessment of solvency”, “Information”, “Output “. The main menu contains information about public housing programs. You can predict solvency by entering the mortgage borrower identification number by clicking the button to predict solvency. The software package is based on processing only real data. The interface of the software complex has such buttons as “Annual income”, “Age of the client”, “Number of children”, “Loan amount”, visualizing data. In addition, there are buttons on the control panel that separate the data into “Training data”, “Test data” and display the results on the screen as a graph.

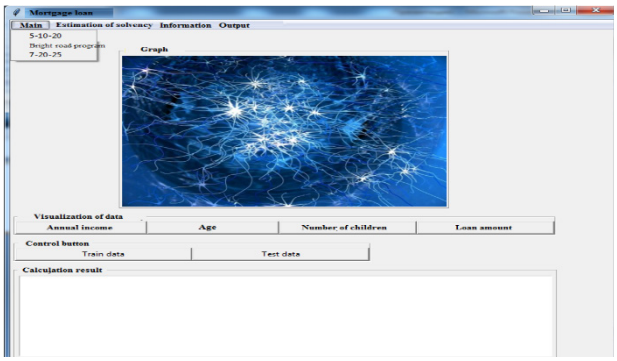


Figure 7 – The interface of the software complex.

MongoDB has proven to be effective at storing data in various formats, that is, unstructured data that can be easily integrated with the established processing system. The amount of real data used for our simulation and system validation in the database was 356,279. The number of their characteristics is about 200. By increasing the number of attributes, we can use them as needed. In the article, I took and edited the attributes I needed. The most necessary data:

- Annual income of the client;
- Amount of children;
- The amount of the loan taken;
- Age of the client.

According to these four features, 1 425 116 records of 356 279 people were trained and processed. In order to accurately determine the client’s solvency, we first train the data and click the “train” button, then, after confirming the accuracy of the trained data, we click the “verify” button to see the processing results.

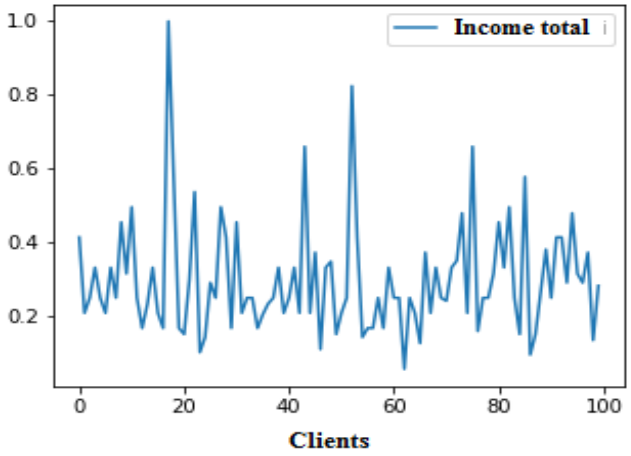


Figure 8 – Customer Forecast for Annual Revenue

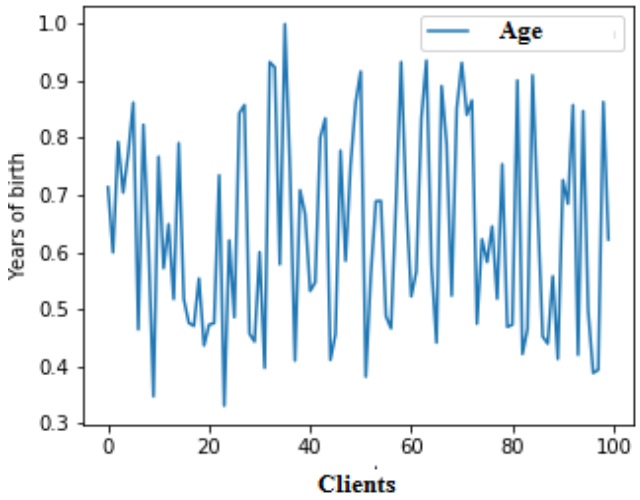


Figure 9 – Customer assumption by age

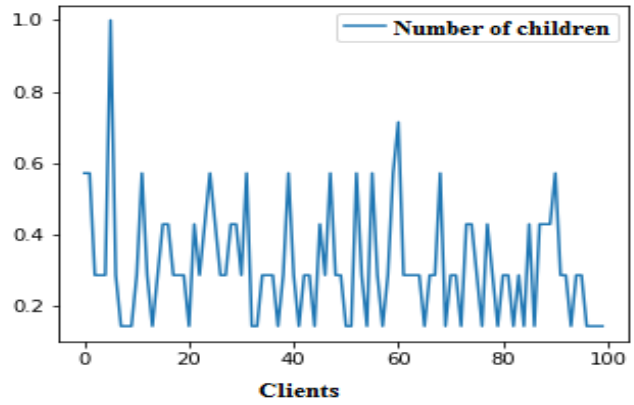


Figure 10 – Make predictions for clients on the number of children they will have

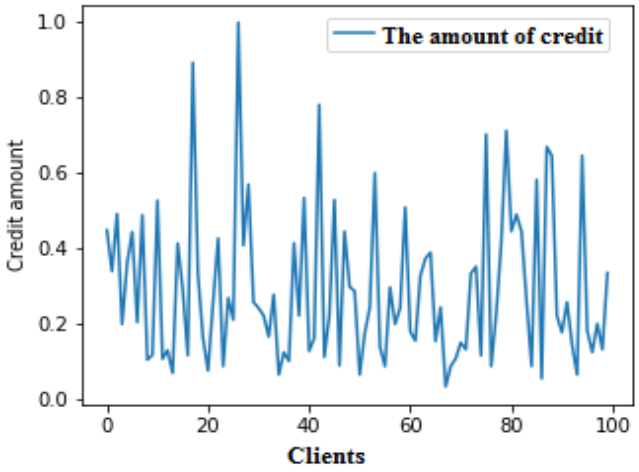


Figure 11 – To make a forecast for customers

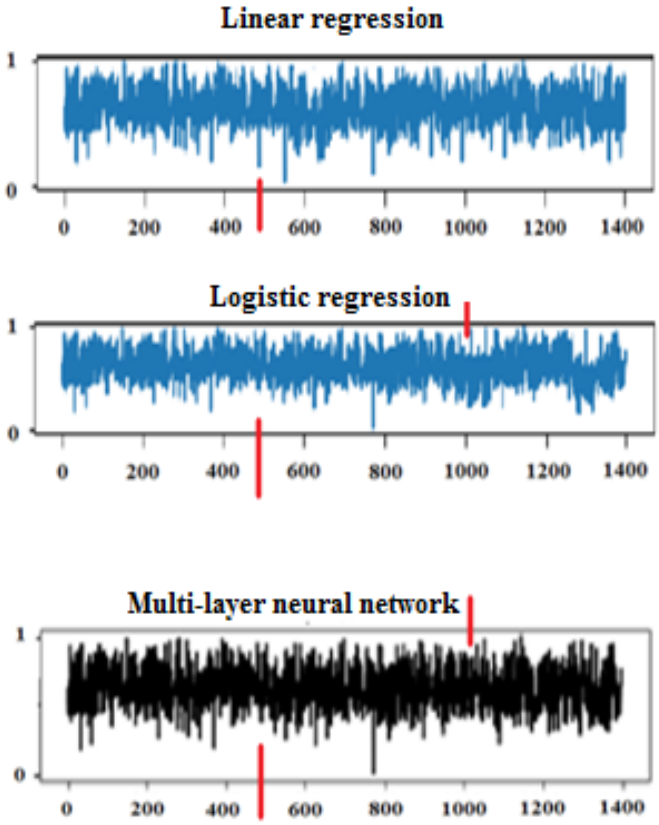


Figure 12 – Slice since data prediction.

On fig. 13-15 shows the results of testing the algorithms. On the graph, the results of the processing algorithms are marked with a red line. Linear regression can handle small data, logistic regression handles medium-sized data, that is, 2 times more data than linear regression, and multilayer neural network shows excellent results when processing large data. As we see in

fig. 13, you can see the differences in the graphs of the three selected algorithms, that is, they have changed due to the increase in the amount of data. This, in turn, is another proof that the use of a multilayer neural network is more efficient when processing BigData.

Factors that indicate the priority in using linear regression are the following:

- The data is used very effectively for linear regression;
- Good results are obtained when processing small amounts of data;
- The theory of linear regression is clear and simple;
- The linear regression method has greatly simplified the work on modeling the system for determining the solvency of mortgage borrowers.

Main disadvantages of linear regression:

- Output parameters, independent variables exceeding the interval [0,1].
- In the obtained results, the values of unknown parameters are often found.

Priorities of the results obtained using the logistic regression method:

- Accurate prediction for medium-sized data, that is, it can be said that the amount of data processed is 2 times more than the data processed using linear regression.

The method of using multilayer neural networks showed very good results in processing large volumes of unstructured data used in the thesis. For the problem being solved, the speed of learning and calculation of neurons shows very good results.

Comparison of methods is shown in the table 3 below.

Below is a graph showing the errors in the results of data processing by three algorithms.

Table 3 – Methods used in data processing and processing results

Method	Data volume	Accuracy (%)	Error (+/-)
Linear regression	$1.43*10^6$	54,62	45,38
Logistic Regression	$1.43*10^6$	68,3	31,7
Neural networks	$1.43*10^6$	78,44	21,56

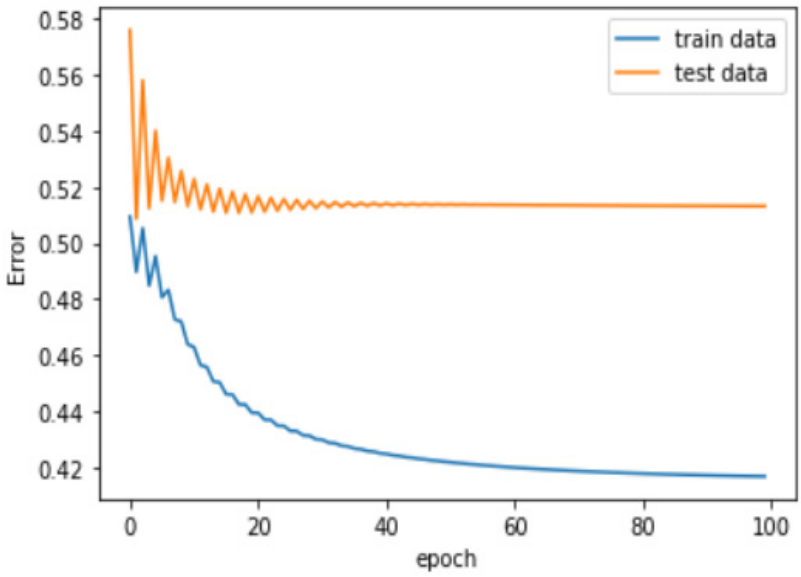


Figure 13 – Error shown by linear regression

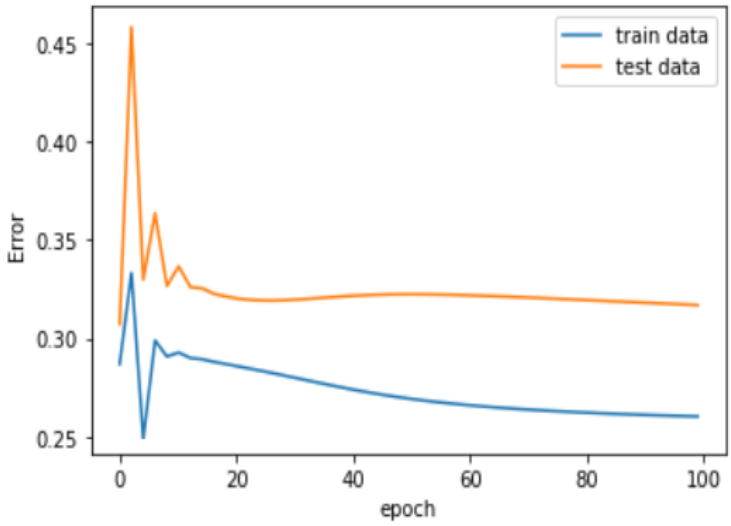


Figure 14 – Error indicated by logistic regression

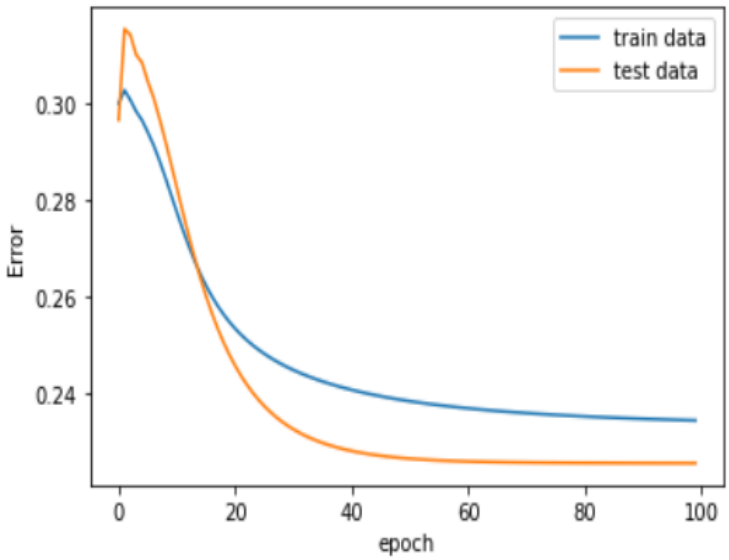


Figure 15 – An error shown by a multi-layer neural network

11. Conclusion

In this article the following tasks have been solved:

1. Methodology and data processing system reviewed.

2. Data mining techniques: Based on linear regression, logistic regression, and multi-layer neural networks, algorithms, and numerical models have been created to process large amounts of data.

3. The quality of the data processing system was assessed and the data was tested.

4. Creating software to predict people's solutions based on big data processing.

The software is implemented according to the following algorithm:

1. Data Preparation:

Convert large unstructured data in CSV format to MongoDB database.

2. The training data will be imported.

3. The imported data is normalized.

4. The processed data was compared with the known response (future matrix, target).

5. Data processed through data mining techniques (linear regression, logistic regression, multi-layer

neural network) to determine the accuracy of the results.

6. If the model encounters too many errors, the loop repeats and continues retrieving data from the MongoDB database until the error rate decreases. This cycle is repeated until the error is reduced and the weighting factors stop changing.

The system created has been successfully tested in the context of mortgage loans provided by financial institutions.

Big Data software helps in determining and predicting a person's solvency. All real data was obtained from financial institutions and the created algorithm proved to be very accurate.

Acknowledgments

This research was funded by the Science Committee of the Ministry of Education and Science of the Republic of Kazakhstan (Grant #AP09259208).

References

1. G.T. Balakayeva, D.K. Darkenbayev, Chris Phillips. Investigation of technologies of processing of Big Data. International Journal of Mathematics and Physics 8, №2, 13(2017).-P.13-18.
2. RubanovV.A. Between management standards and the information element // Technological forecast. – 2010. – No. 3.
3. Big data // https://ru.wikipedia.org. 07/20/2010.
4. Nurlybaeva K.K., Balakayeva G.T. Algorithmization of the process of building scoring models // Bulletin of KazNTU. Series of technical sciences. -2014.-№6(106). – P.195-200.
5. Big Data Big Opportunity //http://www.oracle.com. 01/28/2012.
6. Semenov Yu.A. Large volumes of data (big data) // http://book.itep.ru. 04/21/2013.
7. Artemov S. Big Data: new opportunities for growing business // Jet Infosystems // http://www.pcweek.ru. 08/20/2008.
8. Doug L. 3D Data Management: Controlling Data Volume, Velocity and Variety // Meta Delta. – 2001. – P.949-951.
9. PetteyC., Goasduff L. Gartner Says Solving Big Data Challenge Involves More Than Just Managing Volumes of Data // http://www.gartner.com.27.06.2011.
10. PetukhovD.BigData. Problem and Solutions //http://www.codeinstinct.pro. 08/11/2012.
11. Lockwood G.K. Conceptual Overview of Map/Reduce and Hadoop // http://www.glenklockwood.com. 06/28/2014.
12. Anshina M. Methods of working with big data and their effectiveness // Big Data conference: opportunity or necessity, March 26. – Moscow 2013.
13. Shipunov A. B., Baldin E. M. Data analysis with R // http://www.inp.nsk.su.05.01.2008.
14. Cherkashenko V.N. Risk management of lending to small and medium-sized businesses //http://bankir.ru. 07/31/2012.
15. Voroshilova I. V., Surina I. V. On the issue of improving the mechanism for assessing the creditworthiness of individual borrowers// http://ej.kubagro.ru. 08/03/2005.
16. ObukhovA.In-Memory. Database in RAM //http://ecm-journal.ru.04.08.2014.
17. FranksB. Taming big data: how to extract knowledge from data arrays using deep analytics. – M.: Mann, IvanoviFerber, 2014. -180p.
18. BonczP., Zukowski M., NesN.MonetDB/X100: Hyper-pipelining query execution // Proceedings of conferenceCIDR.-2005.
19. BonczP.A., Kersten M.L. MILprimitives for querying a fragmented world // VLDB Journal. – 1999.-№8(2). -P.101–119.
20. StonebrakerM., BatkinA., ChenX., CherniackM., FerreiraM., LauE., LinA., MaddenS. R., O'NeilE. J., O'NeilP. E., RasinA., TranN., ZdonikS. B.C-Store: A Column-Oriented DBMS // VLDB Journal. – 2005.-P.553-564.
21. AbadiD.J., MaddenS., HachemN. Column Stores vs. Row Stores: How Different Are They Really? // Proceedings of the ACM SIGMOD International Conference on Management of Data. – Vancouver.– 2008.
22. Gavrilov D. Analytical platform – what is it? // http://www.abc.org.ru. 12/28/2006.
23. G. T. Balakayeva, C. Phillips, D. K. Darkenbayev, M. Turdaliyev. Using NoSQL for processing unstructured Big Data. News of the National Academy of sciences of the Republic of Kazakhstan. ISSN 2224-5278 Volume 6, Number 438, 2019.- P.12 – 21.
24. G. Balakayeva, D.Darkenbayev. The solution to the problem of processing BigData using the example of assessing the solvency of borrowers. Journal of Theoretical and Applied Information Technology-2020. Vol.98. No 13. P.-2659-2670. ISSN: 1992-8645.
25. G.Balakayeva et al: Digitalization of enterprise with ensuring stability and reliability. Informatics, Control, Measurement in Economy and Environmental Protection. Vol 13, No 1, 2023, 54-57. http://doi.org/10.35784/iapgos.3295.