IMPROVING BANKING AND FINANCIAL SERVICES IN GHANA WITH BIG DATA ANALYTICS, A CASE STUDY OF AMANTIN AND KASEI COMMUNITY BANK

Robert Gaayire[1], Solomon Nii Nikoi[2], Rufia Adams[3]
[1]Mathematics & ICT Department, Atebubu College of Education, Attebubu – Bono East Region, Ghana
[2]Information Technology Education Dept., AAM University of Skill Training & Entrepreneurial Development (AAMUSTED), Kumasi, Ghana
[1]Mathematics & ICT Department, Al-Faruq College of Education, Wenchi – Bono East Region, Ghana

Authors' contributions

This work was carried out in collaboration among all authors. Authors RGand SNN designed the study and wrote the first draft of the manuscript, and performed the test analysis, managed the analyses of the study. Author RA managed the literature searches and the overall organization of the writeup. All authors read and approved the final manuscript.

ABSTRACT: *The IBM Intelligent Customer Analytics for Recognition and Exploration (iCARE) framework was introduced in order to study banking customer behaviors from banking large data using analytical modeling approaches and techniques tailored to a specific business situation. In the study, the experimental design research technique was employed, which supported the researcher in grasping and integrating behavioral patterns as well as outlining the problem in a framework that attempted to attain a set of goals. The iCARE solution combines IBM software platforms and enormous data processing capability with bespoke data analytical models, giving deeper consumer insights to address the specific business goals and data environment of Amantin and Kasei Community Bank. The study's findings were that Amantin and Kasei Community Bank were able to estimate that clients with disabilities would use IBM's iCARE platform.Amantin and Kasei Community Bank were able to estimate that users with mobile wallet balances of GHc400 or less are likely to drop and design solutions to maintain them, as well as aid them in gaining many new clients, using IBM's iCARE architecture. As a consequence, banks and organizations that deal with vast volumes of data should use iCARE to get insight from the data generated.*

Keywords: *IBM, Intelligent Customer Analytics for Recognition and Exploration (iCARE), Big Data.*

INTRODUCTION

In 1976, the Ghanaian government established Rural Banks to transfer funds to productive rural firms and promote rural development. Amantin & Kasei Community Bank Limited was founded in 1995 and now has seven branches. Ghana, with its commercial technology marketplaces and modernization legacy, is in a great position to become a regional centre for big data and analytics governance. Hadoop, a software tool, is responsible for Big Data, which is a sort of data drift. Big Data Analytics is the use of cutting-edge analytical techniques to massive and complex datasets. Data storage, analysis, visualization, querying, data updating, and privacy are all involved. Before making conclusions, it is critical to comprehend the data's in-depth ideas.

Apache Hadoop is a Big Data solution that is altering people's perceptions of dealing with unstructured data. It is intended to scale from a single machine to a big cluster of computers by offering vast amounts of storage and fast data processing. In order to make decisions and retain clients, Amantin and Kasei Community Rural Bank in Ghana must use big data technology solutions. Technology has advanced tremendously in the last five years, with the introduction of social networks and new information sources creating vast volumes of data in the business environment. Big Data is becoming increasingly important in the financial sector, with banks using mobile banking and loan financing to reach out to clients in rural locations.

To handle Amantin and Kasei Community Bank Limited's data and devise approaches to retain existing clients and recruit new ones, big data technologies must be deployed. This prompted the researcher to investigate how Big Data technology may be utilized to tackle the problem.

This study aims to offer a big data analytics framework for Amantin and Kasei Community Bank Limited, as well as to deploy relevant technology that is suited to their data's scalability. The framework is intended to anticipate customer retention and fraud detection, and its structure is modular, allowing fresh research work to update the content and relevant discovery set of criteria. Using Python, the Hadoop Ecosystem, Data Mining Techniques (SVM and K-Means), and Data Mining Algorithms, the rapid application development (RAD) methodology will be utilized to construct a big data analytics platform for Amantin and Kasei Community Bank (SVM and K-Means). Creating a virtual machine for the Hadoop Ecosystem, cleaning and integrating data, training suggested models with specified techniques, and testing the train model with bank data.

LITERATUREREVIEW

In today's environment, data mining is becoming increasingly crucial, and six powerful data excavation techniques are highlighted. RapidMiner is a java programming tool that uses a template-based framework to perform novel analysis. WEKA is a non-Java tool that is used in a variety of programs and applications to evaluate agricultural data. RapidMiner provides functionality akin to pre-processing, information visualization, predictive analytics, and applied mathematics modeling, analysis, and preparation.

WEKA allows data processing, grouping, stepping back, visualizing, and feature selection. R is a GNU project that was established using R, an open-source programming language and software package for applied math computations and visualization. Python is an opensource program that is strong and suitable for both beginners and professionals. It includes machine learning, bioinformatics extensions, and text mining functionalities[1].

The Konstanz Information Miner (KNIME) is a tool that divides data pre-processing into three categories: removal, transformation, and loading. It is a free and open-source information analysis, coverage, and incorporation medium written in Java and built on Eclipse.

Due to its extensive collection of language manipulation tools written in Python and user customization, NLTK is the greatest language processing tool.

Data mining is used to collect business knowledge and validate valuation, client preferences, product positioning, sales influence, client contentment, and corporate gains/proceeds. It is used as an analytical technique for analyzing information in the healthcare business[2].

According to [3], data mining and computer-aided analysis should be used to broaden the scope and natural environment of data mining in the healthcare industry.

Data mining tools, according to specialists [4], are utilized to regulate the link between dose and probable adverse effects.

When data mining technologies are used in marketing and retail, marketers will be able to identify patterns of buying behavior for goods and services.

It enables merchants to identify relationships between consumer attributes. In this case, it helps to support the business operations for business revenues as well as to purchase the connected goods.

Data mining applications have transformed the education industry, enhancing efficiency, decreasing school dropout rates, and increasing student retention. They play various essential functions in the educational sector, including maximizing system efficiency, lowering drop-out rates, and raising retention rates[2]

Data security is the safeguarding of data against unwanted access and usage. Confidentiality, Integrity, and Availability are abbreviated as CIA[4]. Mandatory Access Control (MAC) was developed in the 1960s to address database security issues by managing separate tuples and data storage to address inquiries about reviewing aggregated data.

[5] provide certain principles to protect remote data privacy and integrity, such as confined baseline processes, request size regulators, perturbation-based procedures, and data interchange. Cloud computing saves consumers money on managing and maintaining hardware infrastructure

Database security technologies are used to ensure that the relational database management system is completely secure[6]. Fortinet FortiDB, IBM Guardium, Imperva Secure Sphere, and McAfee are among them, as are TrustwareDbProtect, HP Security Voltage, Oracle Advanced Safety, Protegrity Data Safety medium, and Vormetric Clear Encoding. Database Action Monitoring (DAM), Database Valuation, and Clear Database Encoding are all safety techniques that may be utilized to secure a database system. Cloud Service Providers (CSPs) provide vast amounts of low-cost storage space.

[6] suggests competent and practical solutions for ensuring data integrity and privacy of outsourced data on cloud servers. They also propose that data in cloud storage be encrypted before it is sent.

Data Safety Design is a layout that assists users in determining when and where to deploy security measures[4]. It is cheap because it is built on relationships between components and how they depend on one another. The architecture's key characteristics are its clarity and uniformity.

Big Data

The Big Data Age is a phenomena in which businesses maximize their information for competitive advantage [7]. This has made Relational Database Management systems difficult to administer, alter, and change to new and continuously changing demands. Big Data is a technology that processes large amounts of high-volume, high-velocity, and high-variety data in order to extract intended data value and assure high veracity[8]. It's popular, although it has no obvious meaning.

Big Dataversus Relational Database

Accordingto[9], Big data aims to overcome the scalability and cost limits that plagued traditional DBMSs.

As a result of the exponential expansion of data in the United States, big data is becoming increasingly significant. It can handle organized, semi-structured, and unstructured data, including real-time data, analytics data, search data, and more. Big Data supports flexible plans as well, whereas DBMS accepts rigid plans[10]. Moreover, Big Data employs a small number of complicated interrelationships, whereas DBMS employs a large number of complex interrelationships[10].

Big Data Technologies

Apache Hadoop is a free open source framework for Big Data applications and features. Other big data technologies covered in this area include Hadoop Distributed File System (HDFS), Map Reduce, and HBase. HDFS manages fault tolerance by replicating various data blocks among data nodes. Map Reduce has user-friendly programming options for fault tolerance, automated parallelization, scalability, and data locality-based optimizations. HBase is a column-oriented database management system that runs on top of HDFS and is designed for sparse data collections. It does not allow relational data storage at all. HBase applications are written in Java and do not contain REST, Avro, or Thrift programming. Every tablemusthave a PrimaryKey, and all access efforts to HBase tables mustuse this Unique Key"[6].

Hive is a component of the Hadoop Ecosystem, which is built on top of HDFS. It enables SQL developers to create Hive Query Language (HQL) reports that are similar to traditional SQL reports. Big-SQL is an IBM SQL interface on a Hadoop-based platform designed to provide database designers with an easy on-ramp for demanding Hadoop data. BigSheet is a simple data collecting tool with a pleasant graphical interface for non-technical business users. It analyzes data in organized and unstructured formats (websites, papers, etc.) and assists company operators in expanding the space of their intelligence data on the web in a timely manner. Users require minimal to no training.

Jaql is a declarative query language designed to minimize parallel processing in huge data collections (Wullianallur & Raghupathi, 2014).

Apache Flink is a distributed and high-performance stream and batch processing system that is open source(Carboneetal.,2015andAlexandrovetal.,2014).Real-time analytics, an infinite data pipeline, historical data processing, and iterative algorithms may all be defined and implemented as pipeline-tolerant data.

According to Zahariaetal (2016),Spark is a free and open source data processing engine that, like Hadoop Map Reduce, distributes data across computers and processes data in parallel.

Apache Hadoop framework and works on low-cost hardware. It assists enterprises in comprehending and analyzing enormous amounts of unstructured data(IBM,2019).

Info Sphere Big Insights is a data analysis and visualization software platform (In for Sphere, 2020). Info Sphere Big Insights is a platform that allows developers, data scientists, and administrators to easily construct and deploy analytics to extract insights from data.

Areas where Big Data can be applied

Big data is used to solve problems in several areas that collect data with Big Data properties. Telecommunications, health, transportation, meteorology, agriculture, government, societal business channels, and other sectors are included. These are a few examples of successful big data implementations[9].

IBM has identified four important areas where Big Data technology may be used to provide smarter and better services. Big data technology may be used to control traffic, learn about customer preferences, identify network bottlenecks, and provide real-time data and phone services. Global Telecom, Ufone, and XO Communications are examples of communication firms that have embraced this technique. To maximize profit and deliver exceptional customer service, a biotechnology business uses sensor data to boost agricultural yield and assess commuters' travel and transit patterns.

[8]suggest that social media is used for the collection of private information and providing good profiled personal services. [11], explained that to manage massive volumes of data generated by customers, Big Data solutions are required.

Automation has resulted in the proliferation of complicated embedded industrial equipment that generates data. Data analysis is utilized to deal with complicated scientific processes, commodities, or facilities, and modern computer-assisted manufacturing need storage and maintenance to maintain efficient quality control[8].Workers and technicians must collaborate to examine data and make educated decisions to maximize products.

Big data technology has been used in healthcare, but it is difficult to use due to its complexity, diversity, and timelines.

Bank

[8] argued that information is an asset for businesses to acquire a competitive advantage. In this age of big data, the amount of data deposited by banks is expanding rapidly, and the type of data is becoming increasingly complex. Banks may increase client experience and income by utilizing predictive analytics and involuntary decision making.

[12],created the Intelligent Consumer Analytics for Recognition and Exploration (iCARE), an architecture that has been given as a way to professionally investigate customer behavior using banking big data and IBM BigInsight.

Big data is used to assist organizations in making tactical decisions based on analysis, such as fraud detection, customer retention, and fraud analytics.

Merits of Big Data Analytics in Banking

Big Data may be used by financial institutions to measure consumer sentiment, enhance service delivery, discover anomalies, manage risk, segment clients, evaluate customer feedback, detect client faith and confidence, and much more. It may be used to monitor consumer usage patterns, create loyalty programs, analyze customer feedback, and identify client faith and trust. It may also be used to detect fraud, manage risk, segment customers, assess customer feedback, detect client faith and trust, and perform other functions.

The Hadoop ecosystem (iCARE BigInsight) is an open source architecture used for data processing and analysis. It is preferred by Amantin and Kasei Comm. Bank due to its ability to tolerate fault and process both structured and unstructured data. It also has a large storage space due to its use of several servers.

Additionally, Hadoop provides better security since it tries to detect fraudulent activities/transaction and helps the bank protect customers' information to maximize profit.

METHODOLOGY

The Design Research Technique helps researchers to comprehend and incorporate behavioral patterns while also defining an issue in a certain setting[13]. It addresses questions like whether big data technology is best for Amantin and Kasei Comm. Rural Bank and which machine learning algorithm suite is employed to assess the bank. Apache Hadoop Ecosystem was the appropriate big data technology.

Design research does not have to result in a solution but should create some good practicable ideas[14].

The Big Data Analytic Framework (BDAF) was created to assist Amantin and Kasei Community Bank Limited in managing and utilizing the huge volumes of data collected every day. The chosen system development method offers quick development because it allows the software process activities of definition, development, and validation to happen concurrently. A high-spec Laptop, Apache Hadoop Environment, Python Jupiter Notebook, and Machine Learning Methods were employed in an Experimental Lab setup.

The Proposed Big Data Architecture

Figure 1 depicts the overall architecture of the proposed solution for Amantin and Kasei Community Bank using big data analytics.Customer retention, new customer acquisition, the suggested analytical model, structured and unstructured data storage, and data processing are all included. Customer retention is entering current customers' data into a system and evaluating it to tell management on which services and goods should be offered to keep them. Making mobile banking and ATM services available to users is part of new customer acquisition, and the suggested analytical model employs support vector machine (SVM) and k-means.

The bank's database stores both structured and unstructured data, with unstructured data being saved in its natural format. Data processing entails the bidirectional processing of all data gathered by the bank.

The Big Data Analyticsat Amantin and Kasei Community Bank Limited

This paper presented a novel method of storing and using data for Amantin and Kasei Community Bank Ltd. Apache Hadoop is the finest architecture for dealing with the exponential rise of massive volumes of data and the rapid processing of complex structured and unstructured data.

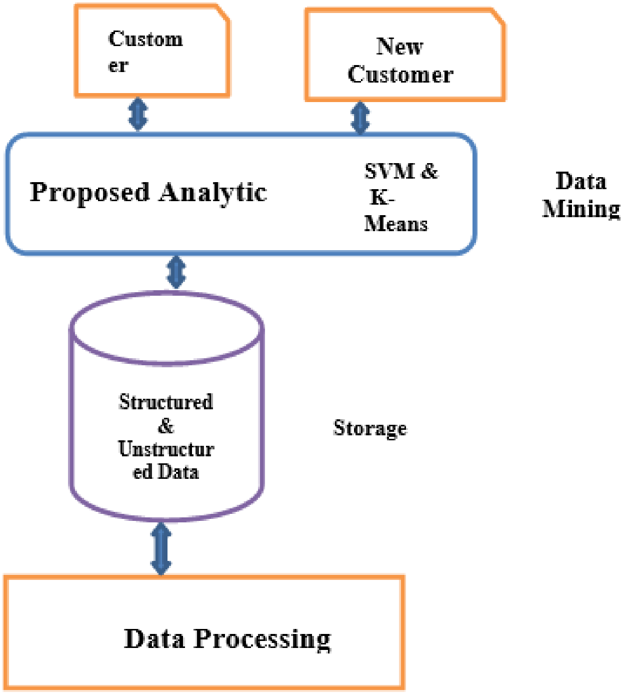


Figure 1: Proposed Big Data architecture

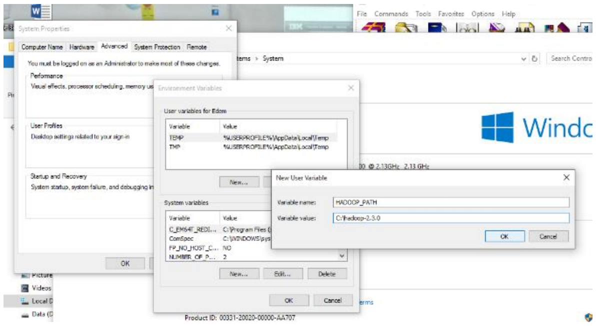
Data processing, cleaning, and integration, design data module, train and test module, interpretation, and prediction are all critical phases in data analysis. Data processing entails acquiring data from multiple sources, cleaning and integrating the data, designing the data module, training and testing the module, and interpreting and predicting the data.

Amantin and Kasei Community Bank Limited possesses a vast quantity of information about its clients' payments and earnings.

Nevertheless, Amantin and Kasei Community Bank Limited want to use Hadoop to analyze this vast amount of data in order to make handling financials for their clients easier and wiser.

Setup Virtual Machine for Hadoop

The first procedure involves downloading Hadoop. In this scenario, Hadoop-3.1.0 was downloaded from the official Hadoop website, extracted, and copied to drive C:. Figure 2 demonstrates how to configure the path and user variables (variable name, value, and path) for Hadoop so that they may be used from the command line.



You must be logged on as an Administrator to make most of these changes.

Visual effects, processor scheduling, memory us

Desktop settings related to your sign-in

System startup, system failure, and debugging in

User variables for Edem

Variable	Value
TEMP	%USERPROFILE%\AppData\Local\Temp
TMP	%USERPROFILE%\AppData\Local\Temp

System variables

Variable	Value
C_EM64T_REDI...	C:\Program Files (
ComSpec	C:\WINDOWS\sys
FP_NO_HOST_C...	NO
NUMBER_OF_P...	2

New User Variable

Variable name: HADOOP_PATH

Variable value: C:\hadoop-2.3.0

Product ID: 00331-20020-00000-AA707

Figure 2 Setting up the user variable for Hadoop

Data Processing

BigIsights is used by acquisition pipeline technology to store streaming data for processing. Data stream management is required to handle data inside the data stream. Amantin and Kasei Comm. Bank data sources included ATMs, interior transactions, internet banking, and mobile banking.

Proprietary API stechnology

Data collecting should necessitate data anonymization and the separation of personal information from commercial activities. Social APIs should be investigated in order to acquire data from many sources utilizing a Massive Parallel Processing SQL engine.

Data Integration

In for Sphere Big Insights (iCARE) offers wrappers/mediators to encapsulate dispersed data as well as automated data and schema mapping technologies, allowing developers to combine data from several perspectives to produce a comprehensive picture.

Data Module

Multi-attribute decision-model technology constructs models from many sources of data using data mining techniques such as Support Vector Machine and K-means clustering.

Model Training and Testing

The validity of the design model was tested through training and testing. Using bank registration forms, similar data from Amantin and Kasei Community Banks was simulated. Each approach was given equal time, and the same amount of data was utilized to train each module. For each module, categories such as Account Name, Account Number, ATM Transactions, Mobile Transactions, Frequent Visit to Bank Hall, and Class (Yes/No) were supplied.

Prediction and Analytics

Amantin and Kasei Community Bank Limited uses model train test results to predict credit card anomaly detection and customer retention. They use Apache Hadoop to manipulate client data collected from many banking products and systems. In the future, the bank will employ Hadoop for risk management in IT and detective work frauds.

Customer Retention

Amantin and Kasei Community Bank Ltd. uses Big Data analytics to control portfolios and create products like mobile banking and ATM machine services. This helps them retain customers and get more new ones. Figure 3 depicts the design process for the proposed framework.

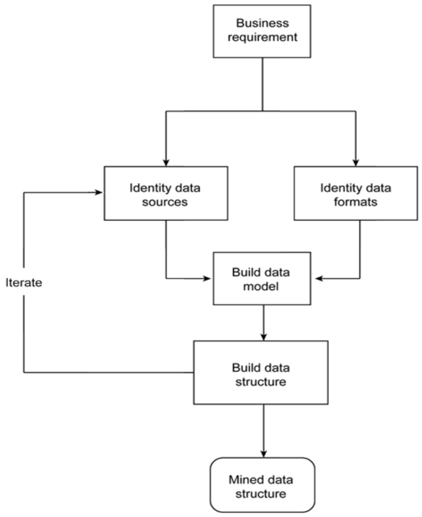


Figure 3: Outline Process for Big Data Analytics Framework

Fraud Detection using Apache Hadoop

Apache Hadoop safeguards unstructured data in the banking business by delivering continuous analysis to strengthen defense and secure assets. Banks' Big Data systems can detect fraud signals, assess them using machine learning, and predict unlawful users and transactions.

Driving value for clients

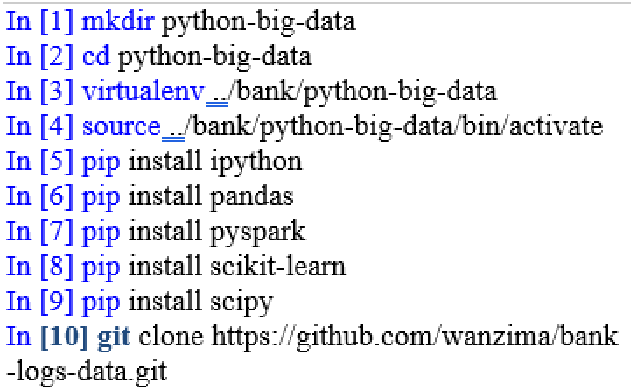
Amantin & Kasei Community Bank provides value to clients by benchmarking their performance against peers and competitors. Hadoop analytics solutions scan internal bank documentation and route it to alternative information sources. Big data analytics is a significant differentiator for the bank in terms of producing affordable, focused investments and personalized customer experiences that aid in company development, increased risk behaviors, and cheaper overall expenses. As the number of consumers grows, the bank will utilize this data to watch consumer activity in real time, giving the exact type of resources needed at any given time.

Data Processing

This process involves finding the relevant data, cleaning it, creating characteristics from the data, and merging data from several databases.

Data Integration

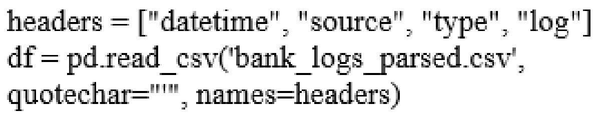
Python libraries install in Jupiter Notebook interactive shell.



```
In [1] mkdir python-big-data
In [2] cd python-big-data
In [3] virtualenv ../bank/python-big-data
In [4] source ../bank/python-big-data/bin/activate
In [5] pip install ipython
In [6] pip install pandas
In [7] pip install pyspark
In [8] pip install scikit-learn
In [9] pip install scipy
In [10] git clone https://github.com/wanzima/bank
-logs-data.git
```

Listing 1: Data integration codes

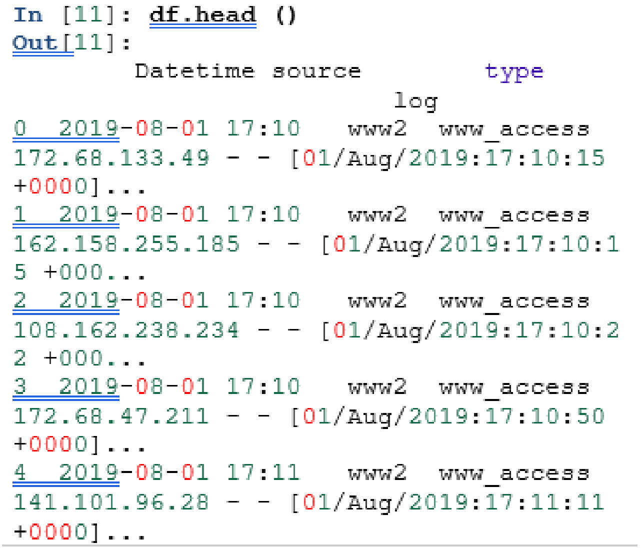
Data operations on our sample bank



```
headers = ["datetime", "source", "type", "log"]
df = pd.read_csv('bank_logs_parsed.csv',
quotechar='"', names=headers)
```

Listing 2: Operation of link to the dataset

Pandas immediately generated a DataFrame object to reflect our CSV file! Examining a sample of the data imported using the head () function.



```
In [11]: df.head ()
Out[11]:
        Datetime source       type
                         log
0  2019-08-01 17:10   www2  www_access
172.68.133.49 - - [01/Aug/2019:17:10:15
+0000]...
1  2019-08-01 17:10   www2  www_access
162.158.255.185 - - [01/Aug/2019:17:10:1
5 +000...
2  2019-08-01 17:10   www2  www_access
108.162.238.234 - - [01/Aug/2019:17:10:2
2 +000...
3  2019-08-01 17:10   www2  www_access
172.68.47.211 - - [01/Aug/2019:17:10:50
+0000]...
4  2019-08-01 17:11   www2  www_access
141.101.96.28 - - [01/Aug/2019:17:11:11
+0000]...
```

Listing 3: Operation of head () function on dataset

Analytical Modelling

This stage focuses on constructing a test design for testing the model, developing models from the dataset, and assessing the constructed model using a data mining method such as Support Vector Machine or K-means.

K-means Model

A parallelized K-means clump, which is an unauthorized machine learning process used to divide n knowledge points into K groups, is the proposed approach for modeling this region. The formula first selects K knowledge points as group centers, then assigns each datum to the closest group and updates the center of each group. Several applications employ the K-means method to decrease complexity and obtain direct data knowledge. It is, nevertheless, prone to errors and variations, making it difficult to describe the commercial pricing of any organization. A bespoke formula is developed as a district of the iCARE solution to improve the toughness of the K-means formula and provide sectional outcomes that are more suitable for a bank.

1. K data was chosen since the following group cores:

a. As the initial group center, use the primary knowledge points.

b. Calculate the shortest spacing between each datum for each defined group center. The Manhattan distance is used. The measure (metric) for two D-dimension knowledge places x and y is defined as

$$d(y, x)= \_(i=1)D|xi-yi |$$

Where xi is the coordinate of x within the ith dimension.

c. At the novel midway, select the information purpose with the most critical least distance from the outline group cores.
d. Repeat steps b. and c. until K group cores are included.
2. Send each datum to the nearest group using the distance measurement calculation in Equation one and the quality K-means approach.
3. Make the following changes to the group midpoints: If there are Jk data points recorded as Xk,1......Xk, jk within the Kth group where k = 1.......K and so the current group midpoint Ck,old, the updated group midpoint is

$$C_{k,new} = \sum_{j=1}^{jk} wk, jXk, j,$$

Where

$$Wk, j = \frac{1}{d(X_{k,j},\ C_{k.old})} / \sum_{j=1}^{jk} \frac{1}{d(X_{k,j},\ C_{k.old})}$$

$$\min \frac{1}{2} w^T w + C \sum \max(0, |yi - (w^T \emptyset(x_n) + b)| \varepsilon)$$

As a consequence, the new group midpoint is the weighted median of each information position during a group, and the corresponding load is compared to the information position gap and the previous group midpoint reciprocally.
4. Distribute the data to their nearest group and eliminate any information xj that is distant from any group that is minimum d

$$(xj, Ck) > 1.1 \leq k \leq K$$

Where r2 is a fixed distance cutoff.
Using the remaining knowledge points in each group, apply two (2) to update group cores.
Recap Steps four and fivetillmax d

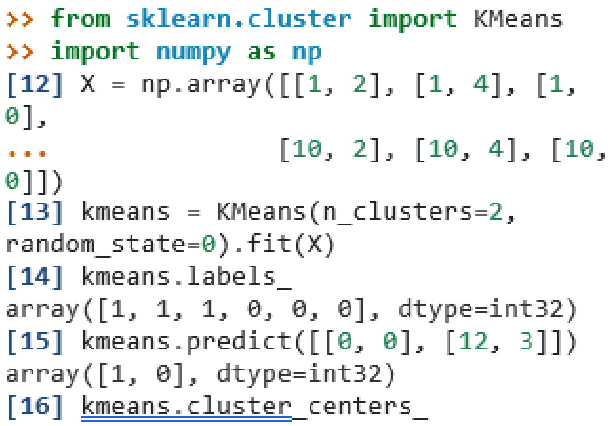
$$(Ck_{,old}, Ck_{,new}) > r2,$$
$$1 \leq k \leq K$$

where r2 could be a pre-planned acceptance threshold.

The MapReduce methodology parallelizes the calculation in each subpart by generating Manhattan spaces between datasets and cluster cores. In parallel, an unfinished weighted add of information points is calculated to bring the group center up to date. Knowledge points far from any center emerge in simultaneously.

4.3.2 Implementation

Since the k-means method is rapid yet slips into local minima, it can be run numerous times to maintain consistent labels and cluster centers. After the last iteration, the estimator will reassign labels to make them consistent with predict on the training set.



```
>> from sklearn.cluster import KMeans
>> import numpy as np
[12] X = np.array([[1, 2], [1, 4], [1,
0],
...                 [10, 2], [10, 4], [10,
0]])
[13] kmeans = KMeans(n_clusters=2,
random_state=0).fit(X)
[14] kmeans.labels_
array([1, 1, 1, 0, 0, 0], dtype=int32)
[15] kmeans.predict([[0, 0], [12, 3]])
array([1, 0], dtype=int32)
[16] kmeans.cluster_centers_
```

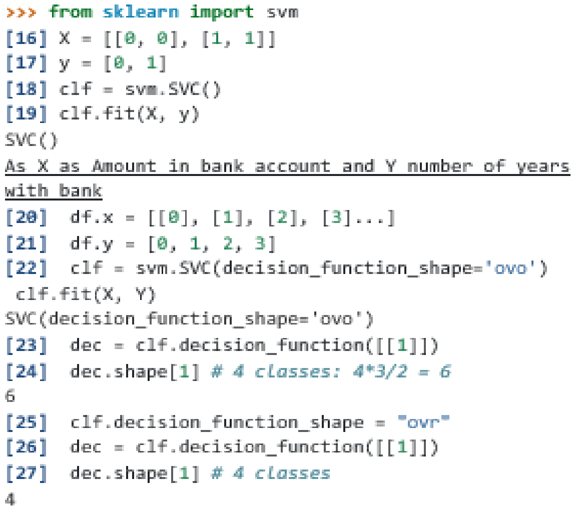
Listing 4: K-means Clustering codes on dataset

Client analysis is used to obtain a better knowledge of clients and their behavior so that their worth may be maximized. To validate the suggested structure, a case study was done using Amantin and Kasei Community Bank Ltd. Five terabytes of data were examined using the iCARE framework to generate insights for retaining active banking clients and identifying customers who were unlikely to abandon supported transactional behavior. To care for the client's energy index, personalized retention methods were established. With their numerous systems, as well as their day-to-day industry and mobile banking, the bank possessed a vast volume of structured data with confusing descriptions.

Support Vector Machine(SVM) Model

A support vector machine constructs a hyperplane or sequence of hyperplanes in a high or infinite dimension space for classification, regression, or other tasks. For classification, the hyperplane with the largest distance to the nearest training data points of any class (so-called functional margin) is the best. The features dual coef_, support vectors_, and intercept_ may be used to identify w Rp and b R so that the prediction produced by sign (wT (x)+b) is correct for the majority of samples.

Linear SVC minimizes hinge loss directly, however it does not contain inner products between samples, therefore the well-known kernel approach cannot be employed. Linear SVC only supports the linear kernel, which is the identity function, because of this. The fundamental challenge is the equation.
SVMs are a type of supervised learning algorithms used for classification, regression, and outlier detection.



```
>>> from sklearn import svm
[16] X = [[0, 0], [1, 1]]
[17] y = [0, 1]
[18] clf = svm.SVC()
[19] clf.fit(X, y)
SVC()
As X as Amount in bank account and Y number of years
with bank
[20]  df.x = [[0], [1], [2], [3]...]
[21]  df.y = [0, 1, 2, 3]
[22]  clf = svm.SVC(decision_function_shape='ovo')
 clf.fit(X, Y)
SVC(decision_function_shape='ovo')
[23]  dec = clf.decision_function([[1]])
[24]  dec.shape[1] # 4 classes: 4*3/2 = 6
6
[25]  clf.decision_function_shape = "ovr"
[26]  dec = clf.decision_function([[1]])
[27]  dec.shape[1] # 4 classes
4
```

Listing 5: Support Vector Machine Clustering codes on dataset

Experimental Result Analysis

Train and Testing Module

A system was created to forecast client retention using the support vector machine and k-means algorithms, with entropy providing the greatest fit.

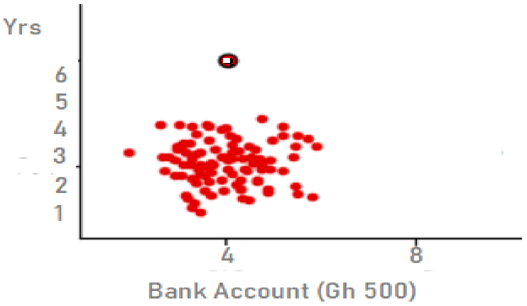


Figure 4.1: K-Means Prediction on Customer Retention

Interpretation of Result

In light of its 86% precision and small number of characteristics, K-means is the best algorithm for predicting customer retention, which influenced the selection of an appropriate big data framework.

Predictive Analytics

The bespoke decision tree is 1.59 times more accurate than the baseline outcome from randomly selected clients, and it may provide recommendations that are easy to grasp.

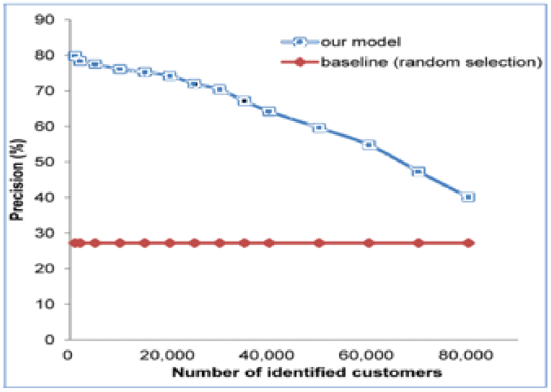


Figure 3: Customer Retention Result

An inventory of customers at high risk of becoming inactive was undertaken, and recommendations were developed to address various perspectives on client behavior. These principles aided the bank's sales staff in sponsoring products or services targeted to the target clientele.

CONCLUSIONS

The article investigated how big data may assist Ghanaian financial institutions by developing an architecture/framework for analytics and decision-making utilizing Apache Hadoop, Hive, HBase, MapReduce, and HDFS.

Predictive models, client retention, actionable information, evidence-based fraud detection, and real-time data gathering and analysis may all be utilized to improve data analytics in Ghanaian financial institutions.

Future Work

Managers and policymakers may utilize Big Data analytics and deep learning techniques to make data useful.

REFERENCES

[1] *S. Singhai and M. Jenna, "A Study on WEKA Tool for Data Preprocessing, Classification and Clustering," 2013, Accessed: Apr. 09, 2023. [Online]. Available: http://www.cs.waikato.ac.nz/~ml/weka/*

[2] *E. Qi, X. Yang, and Z. Wang, "Data mining and visualization of data-driven news in the era of big data," Cluster Comput, vol. 22, no. 4, pp. 10333–10346, Jul. 2019, doi: 10.1007/S10586-017-1348-8/METRICS.*

[3] *M. L. Kolling et al., "Data Mining in Healthcare: Applying Strategic Intelligence Techniques to Depict 25 Years of Research Development," International Journal of Environmental Research and Public Health 2021, Vol. 18, Page 3099, vol. 18, no. 6, p. 3099, Mar. 2021, doi: 10.3390/IJERPH18063099.*

[4] *S. Shukla, J. P. George, K. Tiwari, and J. V. Kureethara, "Data Security," SpringerBriefs in Applied Sciences and Technology, pp. 41–59, 2022, doi: 10.1007/978-981-19-0752-4_3/COVER.*

[5] *G. S. Bahr, L. M. Mayron, and H. J. Gacey, "Cyber risks to secure and private universal access," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 6765 LNCS, no. PART 1, pp. 433–442, 2011, doi: 10.1007/978-3-642-21672-5_47/COVER.*

[6] *H. H. Jaber, "Relational Database Security Enhancements," 2008.*

[7] *T. Davenport, Big Data at Work: Dispelling the Myths, Uncovering the Opportunities - Thomas Davenport - Google Books. 2016. Accessed: Apr. 10, 2023. [Online]. Available: https://books.google.com.gh/books?hl=en&lr=&id=pUO_AgAAQBAJ&oi=fnd&pg=PP1&dq=related:_SJYMBUYLYUJ:scholar.google.com/&ots=UihowKNhvO&sig=5K2fvK48jlxeJyZZBfRjsOFhFN0&redir_esc=y#v=onepage&q&f=false*

[8] Y. Demchenko, ... C. D. L.-2014 I., and undefined 2014, "Defining architecture components of the Big Data Ecosystem," ieeexplore.ieee.org, Accessed: Apr. 10, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6867550/

[9] S. Cheah and S. Wang, "Big data-driven business model innovation by traditional industries in the Chinese economy," Journal of Chinese Economic and Foreign Trade Studies, vol. 10, no. 3, pp. 229–251, 2017, doi: 10.1108/JCEFTS-05-2017-0013/FULL/XML.

[10] Z. Sun, Y. H.-J. of C. I. Systems, and undefined 2021, "The spectrum of big data analytics," Taylor & Francis, vol. 61, no. 2, pp. 154–162, 2019, doi: 10.1080/08874417.2019.1571456.

[11] S. S. Darvazeh, I. R. Vanani, and F. M. Musolu, "Big Data Analytics and Its Applications in Supply Chain Management," 2020. Accessed: Apr. 10, 2023. [Online]. Available: https://library.oapen.org/bitstream/handle/20.500.12657/43835/1/external_content.pdf#page=189

[12] N. Sun, J. G. Morris, J. Xu, X. Zhu, and M. Xie, "ICARE: A framework for big data-based banking customer analytics," IBM J Res Dev, vol. 58, no. 5–6, 2014, doi: 10.1147/JRD.2014.2337118.

[13] D. P. S. Andrew, P. M. Pedersen, and C. D. McEvoy, Research Methods and Design in Sport Management - Damon P.S. Andrew, Paul M. Pedersen, Chad D. McEvoy - Google Books. 2019. Accessed: Apr. 10, 2023. [Online]. Available: https://books.google.com.gh/books?hl=en&lr=&id=9LG8DwAAQBAJ&oi=fnd&pg=PR1&dq=the+design+research+technique+helps+researchers+to+comprehend+and+incorporate+behavioral+patterns+while+also+defining+an+issue+in+a+certain+setting&ots=i0uSTShz6S&sig=4htCisimYGIo0V2hp4TcEeYs1Vs&redir_esc=y#v=onepage&q&f=false

[14] C. Williams, "Research Methods," Journal of Business & Economics Research (JBER), vol. 5, no. 3, p. 65, Mar. 2007, doi: 10.19030/JBER.V5I3.2532.